OPTIMUM QUANTIZATION

JAMES D. BRUCE

TECHNICAL REPORT 429

MARCH 1, 1965

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
RESEARCH LABORATORY OF ELECTRONICS
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory in which faculty members and graduate students from numerous academic departments conduct research.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, by the JOINT SERVICES ELECTRONICS PROGRAMS (U. S. Army, U. S. Navy, and U. S. Air Force) under Contract No. DA36-039-AMC-03200(E); additional support was received from the National Science Foundation (Grant GP-2495), the National Institutes of Health (Grant MH-04737-04), and the National Aeronautics and Space Administration (Grant NsG-496).

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 429

March 1, 1965

OPTIMUM QUANTIZATION

James D. Bruce

This report is based on a thesis submitted to the Department of Electrical Engineering, M.I.T., May 15, 1964, in partial fulfillment of the requirements for the degree of Doctor of Science.

(Manuscript received November 11, 1964)

Abstract

The exact expression for the quantization error as a function of the parameters defining the quantizer, the error-weighting function, and the amplitude probability density of the quantizer-input signal is presented. An algorithm is developed that permits us to determine the specific values of the quantizer parameters that define the optimum quantizer. This algorithm is then extended so that optimum quantizers can be determined for the case in which the quantizer-input signal is a message signal contaminated by noise. In each of these cases the algorithm is based on a modified form of dynamic programming and is valid for both convex and nonconvex error-weighting functions. Examples of optimum quantizers designed with the first of these two algorithms for a representative speech sample are presented. The performance of these optimum quantizers is compared with that of the uniform quantizers.

Author

TABLE OF CONTENTS

I. INTRODUCTION TO QUANTIZATION

Quantization is the nonlinear, zero-memory operation of converting a continuous signal into a discrete signal that assumes only a finite number of levels (N). Quantization occurs whenever physical quantities are represented numerically. In quantization the

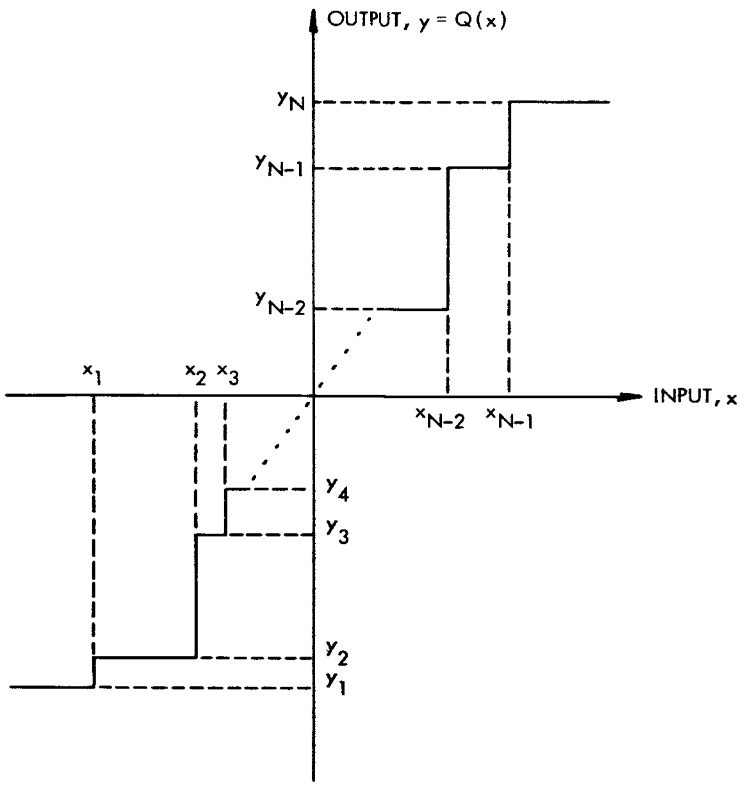

Fig. 1. Input-output relationship of the N-level quantizer.

primary objective is faithful reproduction of the quantizer-input signal at the quantizer output terminals. Such reproduction will be subject to some fidelity criterion such as minimum mean-square error between the quantizer-input signal and its corresponding output. Figure 1 illustrates the input-output characteristic of a N-level quantizer.

1.1 HISTORY OF QUANTIZATION

W. F. Sheppard is the first person who studied a system of quantization. In 1898, he published a paper[1] indicating a method by which the most probable values of the moments of a table of values can be determined from calculations on the members of the table rounded off to points equidistant on a scale. This rounding-off operation is equivalent to uniform quantization or, as it is usually called, analog-to-digital conversion. The input-output characteristic of an analog-to-digital converter is shown in Fig. 2.

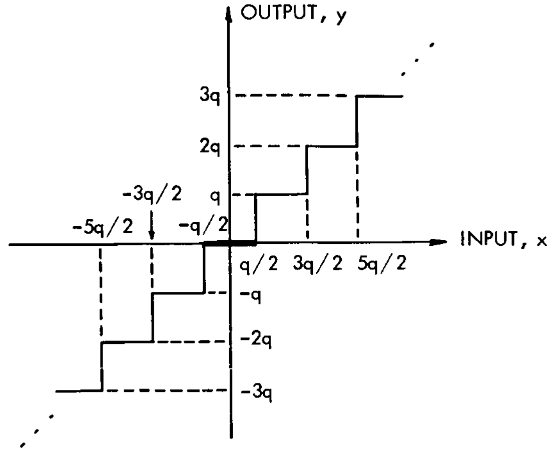

Fig. 2. Input-output characteristic of the analog-to-digital converter.

In their separate investigations of analog-to-digital conversion Widrow[2] and Kosyakin[3] have been able to show that if the characteristic function corresponding to the amplitude probability density of the quantizer-input signal is identically zero outside of some band and if the converter step size "q" is smaller than some critical value related to this bandwidth, then the amplitude probability density of the error signal, the difference between the analog-to-digital converter's input and output signals, will be given by

$$p_e(\lambda) = \begin{cases} 1/q & -q/2 \leq \lambda \leq q/2 \\ 0 & \text{elsewhere} \end{cases} .$$

This density is pictured in Fig. 3.

With the advent of pulse code modulation[4] studies were initiated concerning the application of this modulation scheme which involves sampling and quantization to the transmission of telephone signals. One of the first investigators was Bennett.[5, 6] In 1948, he analyzed the power density spectrum of the analog-to-digital converter's error signal.

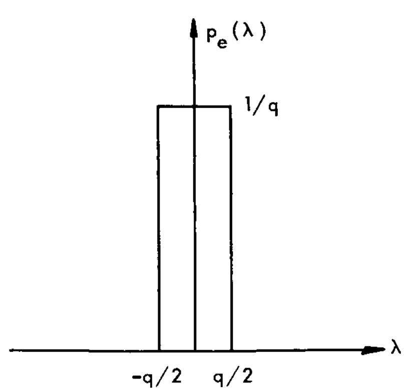

Fig. 3. Analog-to-digital conversion error probability density as derived by Widrow.

Similar studies have been performed by Velichkin[7] and Ruchkin[8] for the quantizer. By assuming that the converter-input signal was a joint Gaussian process with a flat, band-limited power density spectrum and that the converter contains "more than a few steps," Bennett was able to demonstrate that the conversion noise was uniformly distributed throughout the signal band. Other phases of his investigations led Bennett to conclude that, in the case of speech, it is advantageous to taper the steps of the quantizer in such a way that finer steps would be available for weak signals. This implies that for a given number of steps coarser quantization occurs near the peaks of large signals. Tapered quantization is equivalent to inserting complementary nonlinear, zero-memory transducers in the signal path before and after an analog-to-digital converter. Figure 4 pictures this system of quantization which is sometimes called companding.[9]

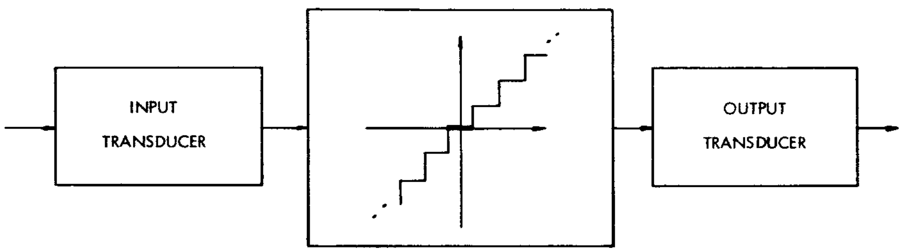

Fig. 4. Model of a tapered quantizer.

Smith,[9] using the model of Fig. 4 for the tapered quantizer, derived the input-output characteristic for the optimum input transducer with respect to the mean-square-error criterion. In doing this, he assumed that the analog-to-digital converter steps were sufficiently small and therefore numerous enough to justify the assumption that the input signal's amplitude probability density is effectively constant within each step, although it varies from step to step. This characteristic has also been obtained by Lozovoy[10] for slightly more general conditions. Several forms of nonoptimum companding have also been investigated and reported.[11-15]

Recently, work in the field of quantization has proceeded basically in two directions. A number of investigators assume that analog-to-digital conversion takes place and attempt to reduce the error by various forms of optimum operation on the converter input and output signals. For example, Katzenelson,[16] Ruchkin,[8] and Stiffler[17] have developed postconverter filters, Graham,[18] working with television signals, has developed preconverter and postconverter filters, Spang[19,20] has developed a linear feedback filter for use around the analog-to-digital converter, and Kimme and Kuo[21] have developed a filter system (see Fig. 5) based on a patent of Cutler.[22] Furman[23,24] and Roberts[25] have both approached the problem in a slightly different manner. Furman has studied the effect of dithering on the analog-to-digital conversion process, while Roberts has applied a similar technique – that of adding pseudo-random noise before conversion and subtracting the same noise after conversion – to the analog-to-digital conversion of television signals.

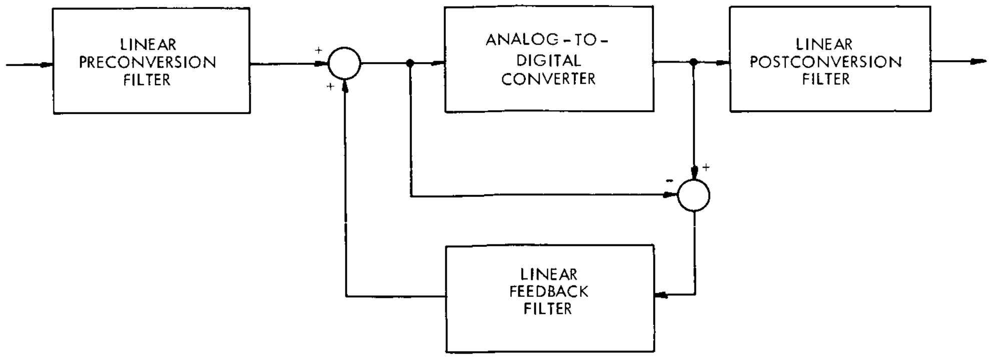

Fig. 5. Quantization system of Kimme and Kuo.

Other investigators such as Max,[26] Lloyd,[27] Gamash,[28] Bluestein,[29] Tou,[30] and Roe[31] have attacked the fundamental problem of designing the optimum quantizer of a specified form. Although some of these investigators have assumed different forms for the allowable class of quantizers, in each case their efforts, for the most part, have been concentrated on specific convex error criteria. Essentially, their approach is to locate all relative extrema of the error surface and select the parameters that define the relative extremum with smallest error as the defining parameters for the optimum quantizer.

Recently, interest has developed in the problem of optimally quantizing signals consisting of the sum of a message signal and an independent noise signal. Myers[32] has studied the amplitude probability density of the error signal when the quantizer-input signal consists of a message signal with flat amplitude density and additive independent Gaussian noise. Stiglitz[33] has determined approximately optimal quantizers for input signals when both the message and the noise are Gaussian processes and the input signal's signal-to-noise ratio is small. Bluestein[29, 34] has shown that if the input signal is composed of a message signal that is constrained to assume only the set of discrete values (y_i) $i = 1, 2, \ldots, N$ plus an independent noise, then the optimum zero-memory filter (minimum mean-absolute-error criterion) will be a quantizer with output levels equal to the set (y_i), $i = 1, 2, \ldots, N$. He also determined asymptotically optimum mean-absolute-error quantizers for the case in which the message signal is continuous.

This problem of determining the optimum quantizer for a signal consisting of a message signal contaminated by additive, independent noise has also been considered by Kuperman.[35] Kuperman makes use of decision-theory concepts to determine the minimum mean-square-error quantizer subject to the constraint that the possible quantizer outputs be uniformly spaced and under the assumption that the quantizer has sufficient levels to justify the assumption that the input-signal amplitude probability density is effectively constant within each level, although changing from level to level.

1.2 BRIEF STATEMENT OF THE PROBLEM

In this report we are primarily concerned with the design of optimum quantizers. We are interested in two cases: first, in which the quantizer-input signal is a message signal; and second, in which the quantizer-input signal is a message signal contaminated (not necessarily additive contamination) by noise. This noise may or may not be statistically independent of the message.

In each of these cases the quantizer will be viewed as a nonlinear, zero-memory filter. Our objective is to develop an algorithm that can be used to determine the quantizer that minimizes some measure of the error, that is, the difference between the message signal and the quantizer-output signal. The measure of the error is taken to be the expected value of some function, called the error-weighting function, of the error. In general, we shall assume that this function is neither symmetric nor convex.

II. QUANTIZATION OF A MESSAGE SIGNAL

2.1 FORMULATION OF THE QUANTIZATION PROBLEM

We have defined quantization as the nonlinear, zero-memory operation of converting a continuous signal into a discrete signal that assumes a finite number of levels (N). The quantizer's input-output characteristic is shown in Fig. 1. We see that the output is y_k when the input signal x is in the range $x_{k-1} \leq x < x_k$. The x_k are called the transition values; that is, x_k is the value of the input variable at which there is a transition in the output from y_k to y_{k+1}. The y_k are called the representation values.

In most communication systems it is desired that the quantized signal be an instantaneous replica of the input message signal. Therefore, the quantizer's desired output is its input signal. Now, in specifying a desired output, we acknowledge that we demand more than the quantizer can accomplish. There will be an error that will be denoted

$$e = x - Q[x]. \tag{1}$$

An appropriate mean value of e will be taken as a measure of how well the quantizer performs with respect to the demands. This measure of the error is given by

$$\mathscr{E} = \int_{-\infty}^{\infty} g[\xi - Q(\xi)]\, p_x(\xi)\, d\xi. \tag{2}$$

Here, $p_x(\xi)$ is the amplitude probability density of the quantizer-input signal x, and $g[\xi - Q(\xi)]$ is a function of the error that we call the error-weighting function. No restrictions are placed on g(e) or $p_x(\xi)$, although usually g(e) is taken to be a non-negative function of its argument because, in general, it is not desirable for positive and negative instantaneous errors to cancel each other.

In order to connect the parameters of the quantizer with the error (we call the measure of the error $\mathscr{E}$ simply the error when there is no chance for confusion), we introduce into (2) the explicit expression for the characteristic of the quantizer,

$$Q(\xi) = y_k \qquad x_{k-1} \leq \xi < x_k \qquad k = 1, 2, \dots, N. \tag{3}$$

Thus we obtain for the error

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} g[\xi - y_{i+1}]\, p_x(\xi)\, d\xi. \tag{4}$$

By definition, x_o will be equal to X_ℓ, the greatest lower bound to the input signal, and x_N will be equal to X_u, the least upper bound to the input signal. Therefore x_o and x_N are constants for any input signal.

From Eq. 4 it is clear that the error $\mathscr{E}$ is a function of the quantizer parameters $(x_1, x_2, \dots, x_{N-1};\ y_1, y_2, \dots, y_N)$; that is,

$$\mathscr{E} = \mathscr{E}(x_1, x_2, \ldots, x_{N-1}; y_1, y_2, \ldots, y_N). \tag{5}$$

The problem before us is to determine the particular x_i $(i = 1, 2, \ldots, N-1)$ and y_j $(j = 1, 2, \ldots, N)$, the quantities that we call X_i and Y_j, which minimize the error $\mathscr{E}$, Eq. 4. Such a minimization is subject to the constraints

$$\begin{aligned} X_\ell = x_o &\le x_1 \\ x_1 &\le x_2 \\ x_2 &\le x_3 \\ &\vdots \\ x_{N-2} &\le x_{N-1} \\ x_{N-1} &\le x_N = X_u \end{aligned} \tag{6}$$

which are explicit in Fig. 1 and Eq. 3. These constraints restrict the region of variation along the error surface $\mathscr{E}$ to a region of that surface so that every point in the region defines a quantizer characteristic $Q(x)$ that is a single-valued function of the input signal x. The error surface is defined on the $(2N-1)$ space specified by considering the $(2N-1)$ quantizer parameters as variables. Such a set of constraints is necessary if the quantizer-input signal is to specify the quantizer-output signal uniquely.

The problem of determining the optimum quantizer then is equivalent to the problem of determining the coordinates of the absolute minimum of the error surface defined by (4) within the region of variation specified by (6).

2.2 DETERMINING THE OPTIMUM QUANTIZER

We have indicated that the problem of designing the optimum quantizer is equivalent to the problem of determining the coordinates of the absolute minimum of the error surface within the region of variation. We know (see, for example, Apostol[36]) that the absolute minimum of the error surface will be either within the region of variation, and therefore at a relative minimum of the error surface, or on the boundary that defines the region of variation. Therefore, given an arbitrary input signal and an arbitrary error-weighting function, we do not know whether the absolute minimum is at a relative minimum or on the boundary. This implies that the technique for determining the optimum quantizer should be a technique that searches for the absolute minimum of the error surface within (or on the boundary of) the region of variation, rather than searching for relative extrema. The method of dynamic programming is such a technique.[36-39] A discussion of the mechanics of dynamic programming is presented in Appendix A.

In order to apply the technique of dynamic programming to the problem of selecting the optimum quantizer (that is, to finding the absolute minimum of the error surface within the region of variation), it is necessary to define three sets of functionals: the error functionals, $\{\epsilon_i(x_i)\}$; the transition-value decision functionals, $\{X_i(x)\}$; and the representation-value decision functionals, $\{Y_i(x)\}$. Each set of functionals has members for i = 1, 2, ..., N. These three sets of functionals are defined in the following manner:

$$
\left.
\begin{aligned}
\epsilon_1(x_1) &= \min_{\substack{y_1 \\ X_\ell = x_o \le x_1 \le X_u}} \left\{ \int_{x_o}^{x_1} d\xi [g(\xi - y_1) p_x(\xi)] \right\} \\
\epsilon_2(x_2) &= \min_{\substack{x_1, y_2 \\ X_\ell \le x_1 \le x_2 \le X_u}} \left\{ \epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi [g(\xi - y_2) p_x(\xi)] \right\} \\
&\vdots \\
\epsilon_i(x_i) &= \min_{\substack{x_{i-1}, y_i \\ X_\ell \le x_{i-1} \le x_i \le X_u}} \left\{ \epsilon_{i-1}(x_{i-1}) + \int_{x_{i-1}}^{x_i} d\xi [g(\xi - y_i) p_x(\xi)] \right\} \\
&\vdots \\
\epsilon_N(x_N) &= \min_{\substack{x_{N-1}, y_N \\ X_\ell \le x_{N-1} \le x_N \le X_u}} \left\{ \epsilon_{N-1}(x_{N-1}) + \int_{x_{N-1}}^{x_N} d\xi [g(\xi - y_N) p_x(\xi)] \right\}
\end{aligned}
\right\} \qquad (7)
$$

$$
\left.
\begin{aligned}
& X_1(x) = X_\ell, \text{ a constant;} \\
& X_2(x) = \text{the value of } x_1 \text{ in the coordinate pair } (x_1, y_2) \text{ that minimizes} \\
& \qquad \left\{ \epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi [g(\xi - y_2) p_x(\xi)] \right\}, \qquad x_2 = x; \\
& \vdots \\
& X_N(x) = \text{the value of } x_{N-1} \text{ in the coordinate pair } (x_{N-1}, y_N) \text{ that minimizes} \\
& \qquad \left\{ \epsilon_{N-1}(x_{N-1}) + \int_{x_{N-1}}^{x_N} d\xi [g(\xi - y_N) p_x(\xi)] \right\}, \qquad x_N = x.
\end{aligned}
\right\} \qquad (8)
$$

$$
\left.
\begin{aligned}
& Y_1(x) = \text{the value of } y_1 \text{ that minimizes} \\
& \qquad \left\{ \int_{x_o}^{x_1} d\xi [g(\xi - y_1) p_x(\xi)] \right\}, \qquad x_1 = x; \\
& Y_2(x) = \text{the value of } y_2 \text{ in the coordinate pair } (x_1, y_2) \text{ that minimizes} \\
& \qquad \left\{ \epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi [g(\xi - y_2) p_x(\xi)] \right\}, \qquad x_2 = x; \\
& \vdots \\
& Y_N(x) = \text{the value of } y_N \text{ in the coordinate pair } (x_{N-1}, y_N) \text{ that minimizes} \\
& \qquad \left\{ \epsilon_{N-1}(x_{N-1}) + \int_{x_{N-1}}^{x_N} d\xi [g(\xi - y_N) p_x(\xi)] \right\}, \qquad x_{N-1} = x.
\end{aligned}
\right\} \qquad (9)
$$

Consider these three sets of functionals. The key to our understanding of their meaning lies in understanding the meaning of the separate members of the error functionals (7). The first member of (7) states that for a given range of the input signal x, or equivalently for a given region of the amplitude probability density of x, $p_x(\xi)$, specified by the boundaries x_o and x_1, we determine the y_1 that minimizes the integral

$$\int_{x_o}^{x_1} d\xi[g(\xi-y_1)p_x(\xi)]. \tag{10}$$

The mechanics of determining this specific y_1 is discussed below. This y_1 is recorded as $Y_1(x)$, $x = x_1$ and the value of the integral (10) for this value of y_1 is recorded as $\epsilon_1(x_1)$. This is done for all x_1 in the range $X_\ell \leqslant x_1 \leqslant X_u$. Thus, if we specify a particular x_1, say, $x_1 = a$, we know that the optimum choice for y_1 is $Y_1(a)$.

Now consider the second member of Eq. 7. This functional indicates that we are considering the quantization of the signal in the input interval $x_o \leqslant x \leqslant x_2$, for a variable x_2, into two levels. In order to perform this operation in the optimum manner, we must minimize the quantity

$$\int_{x_o}^{x_1} d\xi[g(\xi-y_1)p_x(\xi)] + \int_{x_1}^{x_2} d\xi[g(\xi-y_2)p_x(\xi)] \tag{11}$$

with respect to the three variables x_1, y_1, and y_2. The first of these two integrals when minimized with respect to y_1 (and it only contains y_1) is simply the first error functional, $\epsilon_1(x_1)$. Then, for a given x_2, we must determine the x_1 and the y_2 minimizing the function

$$\epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi[g(\xi-y_2)p_x(\xi)]. \tag{12}$$

The x_1 that minimizes (12) is recorded as $X_2(x)$, $x = x_2$; the y_2 that minimizes the expression is recorded as $Y_2(x)$, $x = x_2$. The value of the expression is recorded as $\epsilon_2(x_2)$. These operations are performed for all x_2 in the range $X_\ell \leqslant x_2 \leqslant X_u$. Therefore, if the region $x_o \leqslant x \leqslant x_2$ is to be quantized into two levels, we know from the decision functionals that the optimum transition value is specified by $X_1 = X_2(x_2)$ and that the two optimum representation values are given by $Y_2 = Y_2(x_2)$ and $Y_1 = Y_1(X_1)$.

Clearly, discussion of this type can be presented for each of the members of (7). Instead of considering every member in turn, let us skip to the last member of (7). Here, we are given the input range $x_o \leqslant x \leqslant x_N$; a variable x_N is assumed. We want to quantize this range into N levels in the optimum manner. This requires that we minimize the quantity

$$\int_{x_o}^{x_1} d\xi[g(\xi-y_1)p_x(\xi)] + \int_{x_1}^{x_2} d\xi[g(\xi-y_2)p_x(\xi)] + \dots + \int_{x_{N-1}}^{x_N} d\xi[g(\xi-y_N)p_x(\xi)] \tag{13}$$

with respect to the parameters $y_1, y_2, \ldots, y_N; x_1, x_2, \ldots, x_{N-1}$. This task is not as difficult as it may seem. Note that the minimum of the first term with respect to y_1 as a function of x_1 is given by the first error functional $\epsilon_1(x_1)$. This is the only term of (13) involving y_1. Thus (13) can be written as the minimization of

$$\epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi[g(\xi-y_2)p_x(\xi)] + \int_{x_2}^{x_3} d\xi[g(\xi-y_3)p_x(\xi)] + \ldots + \int_{x_{N-1}}^{x_N} d\xi[g(\xi-y_N)p_x(\xi)] \tag{14}$$

with respect to $y_2, y_3, \ldots, y_N; x_1, x_2, \ldots, x_{N-1}$. But note that the minimization of the first two terms of Eq. 14 with respect to y_2 and x_1 as a function of x_2 is given by $\epsilon_2(x_2)$. And these are the only terms involving y_2 and x_1. Thus Eq. 14 can be written equivalently as the minimization of

$$\epsilon_2(x_2) + \int_{x_2}^{x_3} d\xi[g(\xi-y_3)p_x(\xi)] + \int_{x_3}^{x_4} d\xi[g(\xi-y_4)p_x(\xi)] + \ldots + \int_{x_{N-1}}^{x_N} d\xi[g(\xi-y_N)p_x(\xi)] \tag{15}$$

with respect to $y_3, y_4, \ldots, y_N; x_2, x_3, \ldots, x_{N-1}$.

This process can be continued until we obtain as an equivalent for (15) the minimization of

$$\epsilon_{N-1}(x_{N-1}) + \int_{x_{N-1}}^{x_N} d\xi[g(\xi-y_N)p_x(\xi)] \tag{16}$$

with respect to x_{N-1} and y_N. For a specific x_N, the x_{N-1} and y_N that minimize (16) are recorded as $X_N(x)$ and $Y_N(x)$, respectively, $x = x_N$. The value of (16) for a specific x_N is recorded as $\epsilon_N(x_N)$. The two decision functionals and the error functional are evaluated for all x_N so that $X_\ell \leq x_N \leq X_u$.

Appendix B gives an alternative presentation of this explanation of the error and decision functionals from the point of view of search paths along the error surface.

Now, we are in a position to use the functionals just derived to determine the parameters defining the optimum quantizer. Note that when $x_N = X_u$ we are considering the entire input-signal range. Thus, $\epsilon_N(X_u)$ is the total quantization error for the optimum N-level quantizer. Then from the definition of $X_N(x)$, the $(N-1)^{th}$ transition value is

$$X_{N-1} = X_N(X_u).$$

Likewise, from the definition of $Y_N(x)$, the $(N)^{th}$ representation value is

$$Y_N = Y_N(X_u).$$

Continuing from our definition of $X_{N-2}(x)$ and $Y_{N-2}(x)$, we find that the $(N-2)^{th}$ transition value is

$$X_{N-2} = X_{N-1}(X_{N-1})$$

and the $(N-1)^{th}$ representation value is

$$Y_{N-1} = Y_{N-1}(X_{N-1}).$$

This process can be continued until finally we have

$$Y_1 = Y_1(X_1),$$

which is the last parameter needed to completely define the optimum quantizer.

2.3 SIMPLIFICATION OF THE ERROR FUNCTIONALS

We have presented a method by which the parameters that define the optimum quantizer can be determined. Now, we want to take a closer look at the error functionals, Eq. 7, with the hope of reducing the complexity of the minimization operation.

We begin by considering the k^{th} member of (7), that is,

$$\epsilon_k(x_k) = \min_{\substack{x_{k-1}, y_k \\ X_\ell \le x_{k-1} \le x_k \le X_u}} \left\{ \epsilon_{k-1}(x_{k-1}) + \int_{x_{k-1}}^{x_k} d\xi [g(\xi - y_k) p_x(\xi)] \right\}. \tag{17}$$

Since $\epsilon_{k-1}(x_{k-1})$ is not a function of y_k, the k^{th} member of (17) may be equivalently written

$$\epsilon_k(x_k) = \min_{\substack{x_{k-1} \\ X_\ell \le x_{k-1} \le x_k \le X_u}} \left\{ \epsilon_{k-1}(x_{k-1}) + \min_{y_k} \int_{x_{k-1}}^{x_k} d\xi [g(\xi - y_k) p_x(\xi)] \right\}. \tag{18}$$

We now limit our consideration to the last term of (18), that is, to

$$\min_{y_k} \int_{x_{k-1}}^{x_k} d\xi [g(\xi - y_k) p_x(\xi)]. \tag{19}$$

From the original statement of the quantization problem, Eqs. 4 and 6, we see that the y_k are unconstrained variables, that is the region of variation for y_k is $-\infty \le y_k \le +\infty$. Therefore, the y_k that minimizes

$$\int_{x_{k-1}}^{x_k} d\xi [g(\xi - y_k) p_x(\xi)] \tag{20}$$

must be a relative extremum of

$$f_k(x_{k-1}, x_k; y_k) = \int_{x_{k-1}}^{x_k} d\xi [g(\xi - y_k) p_x(\xi)] \tag{21}$$

with respect to y_k. If g is a continuous function of its argument, the y_k minimizing (20) will be a solution of

$$0 = \int_{x_{k-1}}^{x_k} d\xi \left\{ p_x(\xi) \left[\frac{\partial}{\partial y_k} g(\xi - y_k) \right] \right\}. \tag{22}$$

In general (22) will have a multiplicity of solutions, each of which will be a function of the remaining parameter of minimization, x_{k-1}. It will be necessary to determine which of these solutions is the one yielding the absolute minimum of (21) by substitution of the solutions in the equation. This particular y_k will be indicated by y_k^*. (It will be shown that if g is strictly convex then (22) will have only one solution.) If g is not continuous, y_k^* will be found by a direct search along the set of possible values for y_k.

Using this result, we may write Eq. 17 as

$$\epsilon_k(x_k) = \min_{\substack{x_{k-1} \\ X_\ell \leq x_{k-1} \leq x_k \leq X_u}} \left\{ \epsilon_{k-1}(x_{k-1}) + \int_{x_{k-1}}^{x_k} d\xi \left[g\left(\xi - y_k^*\right) p_x(\xi) \right] \right\}. \tag{23}$$

This indicates that the number of parameters over which the formal minimization must be performed can be reduced from (2N-1) in Eq. 7 to (N-1) when the error functionals are written in the form of (23):

$$\left.\begin{aligned}
\epsilon_1(x_1) &= \int_{x_o}^{x_1} d\xi \left[g\left(\xi - y_1^*\right) p_x(\xi) \right] \\
&\qquad X_\ell = x_o \leq x_1 \leq X_u \\
\epsilon_2(x_2) &= \min_{\substack{x_1 \\ X_\ell \leq x_1 \leq x_2 \leq X_u}} \left\{ \epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi \left[g\left(\xi - y_2^*\right) p_x(\xi) \right] \right\} \\
&\vdots \\
\epsilon_N(x_N) &= \min_{\substack{x_{N-1} \\ X_\ell \leq x_{N-1} \leq x_N \leq X_u}} \left\{ \epsilon_{N-1}(x_{N-1}) + \int_{x_{N-1}}^{x_N} d\xi \left[g\left(\xi - y_N^*\right) p_x(\xi) \right] \right\}.
\end{aligned}\right\} \tag{24}$$

From a practical point of view we cannot determine the error functionals (or for that matter the decision functionals) in closed form but only at several points that specify a grid. This result, then, enables us to determine each of the error functionals by a one-dimensional search for each value of x_k, rather than by a two-dimensional search, thus substantially reducing the complexity and length of the computations.

2.4 QUANTIZATION – A SECOND POINT OF VIEW

We have defined the quantization error to be

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi [g(\xi - y_{i+1}) p_x(\xi)]. \tag{25}$$

Now, we want to derive an alternative expression for the error which will allow us to interpret the optimum quantizer from a different point of view.

We begin by defining the random variable λ as

$$\lambda = \xi - y_{i+1}. \tag{26}$$

Clearly, λ is a random variable corresponding to the amplitude of the error signal. Upon substitution of (26) in (25) $\mathscr{E}$ becomes

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i - y_{i+1}}^{x_{i+1}-y_{i+1}} d\lambda [g(\lambda) p_x(\lambda + y_{i+1})]. \tag{27}$$

But this equation can be written

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{-\infty}^{\infty} d\lambda [g(\lambda) p_x(\lambda+y_{i+1})] \cdot \{u_{-1}[\lambda-(x_i-y_{i+1})]-u_{-1}[\lambda-(x_{i+1}-y_{i+1})]\}, \tag{28}$$

where $u_{-1}(a)$ is defined by

$$u_{-1}(a) = \begin{cases} 1 & a \geq 0 \\ 0 & a < 0. \end{cases}$$

By interchanging the order of summation and integration, Eq. 28 becomes

$$\mathscr{E} = \int_{-\infty}^{\infty} d\lambda \; g(\lambda) \sum_{i=0}^{N-1} p_x(\lambda+y_{i+1}) \cdot \{u_{-1}[\lambda-(x_i-y_{i+1})]-u_{-1}[\lambda-(x_{i+1}-y_{i+1})]\} \tag{29}$$

We want to identify the term of (29) that involves the summation, that is,

$$\sum_{i=0}^{N-1} p_x(\lambda+y_{i+1})\{u_{-1}[\lambda-(x_i-y_{i+1})]-u_{-1}[\lambda-(x_{i+1}-y_{i+1})]\}. \tag{30}$$

Consider the k^{th} term of this sum. This term represents the portion of the input signal's amplitude probability density lying between $x_{k-1} \leq \xi < x_k$. This has now been shifted so that the representation value y_k corresponding to this interval is at the origin. Thus the term is the contribution to the amplitude probability density of the error by input signals in the range

$$x_{k-1} \leq \xi < x_k.$$

This permits us to conclude that the sum, Eq. 30, is the amplitude probability density of the error,

$$p_e(\lambda) = \sum_{i=0}^{N-1} p_x(\lambda+y_{i+1})\{u_{-1}[\lambda-(x_i-y_{i+1})]-u_{-1}[\lambda-(x_{i+1}-y_{i+1})]\} \tag{31}$$

and therefore that Eq. 29 may be written

$$\mathscr{E} = \int_{-\infty}^{\infty} d\lambda[g(\lambda)p_e(\lambda)]. \tag{32}$$

Now recall that we minimize $\mathscr{E}$ with respect to the x_k and y_j when we design the optimum quantizer. With respect to the error given by (32) these parameters x_k and y_j are involved in the expression for $p_e(\lambda)$. Therefore, we conclude that the problem of designing the optimum quantizer is equivalent to shaping the amplitude probability density of the error signal so that some property of this density specified by the error-weighting function g is minimized. This shaping is constrained by the number of levels permitted in the quantizer (N) and by the input-signal amplitude probability density $p_x(\xi)$.

III. SOME RESULTS FOR RESTRICTED ERROR-WEIGHTING FUNCTIONS

An alogrithm that allows us to determine the parameters that define the optimum quantizer has been developed. We now want to examine this solution for a class of error-weighting functions which we shall call monotonic error-weighting functions. A monotonic error-weighting function $g(e)$ is a function such that for any $e \geq 0$ and any $\delta > 0$, $g(e+\delta) > g(e)$; and for any $e \leq 0$ and any $\delta < 0$, $g(e+\delta) > g(e)$. That is, $g(e)$ is a monotonically decreasing function for negative error and a monotonically increasing function for positive error.

In particular, we are interested in examining the possibility that the absolute minimum will be at a relative minimum of the error surface. This will lead to a discussion of the properties of the relative extremum of the error surface within the region of variation.

3.1 NATURE OF THE ABSOLUTE MINIMA

Our primary concern is to prove that the absolute minimum of the quantization error within the region of variation is at a relative extremum, a minimum, of the error surface, rather than on the boundary defining the region of variation, if the error-weighting function is monotonic and if the quantizer-input signal amplitude probability density is not entirely discrete.

We begin by assuming that the quantizer-input signal x, a signal that is not entirely discrete, is quantized into (N) levels by a quantizer with transition values

$$\{x_i\}, \quad i = 1, 2, \ldots, N-1$$

and representation values

$$\{y_j\}, \quad j = 1, 2, \ldots, N.$$

The quantization error for this set of quantizer parameters is

$$\mathscr{E}_N = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi [g(\xi - y_{i+1}) p_x(\xi)]. \tag{33}$$

We shall construct an (N+1)-level quantizer in the following manner:

1. Select an interval of the N-level quantizer such that the continuous portion of the input amplitude probability density is not zero everywhere in this interval. The transition values at the end of this quantization interval will be labeled x_{k-1} and x_k.

2. Select an $(N)^{th}$ transition value a at a point in this interval where the continuous portion of the density is nonzero and such that $a \neq y_k$.

3. If $a > y_k$ select as the representation value for the interval $x_{k-1} \leq \xi < a$ the value y_k, and for the interval $a \leq \xi < x_k$ the value a. If $a < y_k$ select as representation value

for the interval $x_{k-1} \leq \xi < \alpha$ the value α, and for the interval $\alpha \leq \xi < x_k$ the value y_k.

4. The remaining parameters of the (N+1)-level quantizer are identical to those of the N-level quantizer.

If we denote by $\mathscr{E}_{N+1}$ the error for the (N+1)-level quantizer constructed in this manner, the error difference

$$\Delta\mathscr{E} = \mathscr{E}_N - \mathscr{E}_{N+1}$$

will be

$$\Delta\mathscr{E} = \int_{x_{k-1}}^{x_k} d\xi[g(\xi-y_k)p_x(\xi)] - \int_{x_{k-1}}^{\alpha} d\xi[g(\xi-y_k)p_x(\xi)] - \int_{\alpha}^{x_k} d\xi[g(\xi-\alpha)p_x(\xi)]. \tag{34}$$

In writing Eq. 34 we have assumed $\alpha > y_k$. A parallel expression can be written for the case $\alpha < y_k$.

If we write the first integral of (34) as the sum of two integrals, upon collection of terms we have

$$\Delta\mathscr{E} = \int_{\alpha}^{x_k} d\xi\{[g(\xi-y_k)-g(\xi-\alpha)]\, p_x(\xi)\}. \tag{35}$$

We observe that since $\alpha > y_k$, the quantity $[g(\xi-y_k)-g(\xi-\alpha)]$ will be positive for all values of ξ in the range $\alpha \leq \xi < x_k$ if g is a monotonic error-weighting function. By construction of the quantizer, $p_x(\xi)$ is not zero over the entire interval of integration, Eq. 35, and therefore

$$\Delta\mathscr{E} > 0. \tag{36}$$

(It should be clear that a similar argument can be presented to show that Eq. 36 holds for $\alpha < y_k$.)

It then follows that if g is a monotonic error-weighting function and if $p_x(\xi)$ is not entirely discrete, there exists at least one (N+1)-level quantizer with less error than any N-level quantizer.

In order to use this result, we must consider some properties of boundary solutions. Solutions on the boundary are in part specified by an equation of the form

$$x_j = x_{j+1}$$

which indicates the parameter on the boundary, since the region of variation is defined by

$$\begin{aligned} X_\ell = x_o &\leq x_1 \\ x_1 &\leq x_2 \\ &\vdots \\ x_{N-1} &\leq x_N = X_u. \end{aligned} \tag{37}$$

The property that we wish to note is that if a quantizer with (N) levels is defined by a point on the boundary, its error cannot be less than the error for the optimum (N-1)-level quantizer. This can be easily verified by examination of the equation that defines the quantization error, Eq. 33. Referring to (33), we realize that a solution on the boundary requires one of the terms in the sum to be

$$\int_{x_j}^{x_{j+1}=x_j} d\xi[g(\xi-y_{j+1})p_x(\xi)].$$

It is clear that the numerical value of such a term is zero. Thus, this term has the effect of reducing the number of effective quantizer levels to (N-1). Therefore, the smallest possible value for the error in this N-level quantizer is the error for the optimum (N-1)-level quantizer.

Now, returning to the particular problem at hand, we recall that we are able to construct at least one (N+1)-level quantizer with less error than any N-level quantizer when the error-weighting function is monotonic and $p_x(\xi)$ is not entirely discrete. Now, since the error is less for at least one (N+1)-level quantizer, the optimum (N+1)-level quantizer must be defined by a relative minimum of the error surface within the region of variation, rather than by a point on the boundary. A solution on the boundary is prohibited by the decrease in error. Since this result is independent of N, we conclude that if g is a monotonic error-weighting function and $p_x(\xi)$ is not entirely discrete, then the optimum quantizer is always defined by a relative minimum of the error surface within the region of variation.

3.2 LOCATION OF THE RELATIVE EXTREMA

The preceding section suggests that it will be of value to locate the relative extrema (and in particular the relative minima) of the error surface as an alternative method of specifying the parameters defining the optimum quantizer. From calculus[40] we know that the quantizer error surface $\mathscr{E}$ will attain a relative extremum (or a saddle point) for those values of the (2N-1)-quantizer parameters that force the (2N-1) first partial derivatives of $\mathscr{E}$ to become zero simultaneously. That is, the surface's relative extrema are solutions of the set of simultaneous equations

$$\begin{aligned} \frac{\partial \mathscr{E}}{\partial y_k} &= 0 \qquad k = 1, 2, \dots, N \\ \frac{\partial \mathscr{E}}{\partial x_i} &= 0 \qquad i = 1, 2, \dots, N-1. \end{aligned} \tag{38}$$

Substituting Eq. 4 in Eq. 38 we have

$$\frac{\partial \mathscr{E}}{\partial y_k} = \int_{x_{k-1}}^{x_k} d\xi \left\{ p_x(\xi) \frac{\partial}{\partial y_k} [g(\xi-y_k)] \right\} = 0, \qquad k = 1, 2, \dots, N \tag{39}$$

$$\frac{\partial \mathscr{E}}{\partial x_i} = [g(x_i-y_i)-g(x_i-y_{i+1})]\, p_x(x_i) = 0, \qquad i = 1, 2, \ldots, N-1. \tag{40}$$

If $p_x(\xi)$ is nonzero between its greatest lower bound X_ℓ and its least upper bound X_u, Eq. 40 becomes

$$g(x_i-y_i) - g(x_i-y_{i+1}) = 0, \qquad i = 1, 2, \ldots, N-1. \tag{41}$$

Therefore, for the case of nonzero $p_x(\xi)$ in the interval $X_\ell < \xi < X_u$, the relative extrema are solutions of (39) and (41). In writing Eqs. 39-41 our only assumption concerning g is that its first derivative exists; that is, g is continuous.

It is of interest at this point in our discussion to observe that if g is a symmetric, monotonic error-weighting function, then (41) may be equivalently written

$$x_i = \frac{y_{i+1} + y_i}{2}, \qquad i = 1, 2, \ldots, N-1. \tag{42}$$

This follows from graphical examination of Eq. 41 and the realization that for symmetric, monotonic error-weighting functions this equation will always have a unique solution.

Joel Max[26] has developed an algorithm to determine the relative extrema of $\mathscr{E}$, using Eqs. 39 and 41 for the special case in which the error-weighting function is $(\cdot)^2$. That is, with $\mathscr{E}$ given by

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi[(\xi-y_{i+1})^2 p_x(\xi)]. \tag{43}$$

This algorithm consists of choosing a value for y_1 and then alternatively applying Eq. 39, which in this case reduces to

$$y_k = \frac{\int_{x_{k-1}}^{x_k} d\xi[\xi p_x(\xi)]}{\int_{x_{k-1}}^{x_k} d\xi[p_x(\xi)]}, \tag{44}$$

and Eq. 42 to determine first approximations to the values of the quantizer parameters that define relative extrema of the error surface. (The first application of (44) will yield x_1, as x_0 and y_1 are known; the first application of (42) will yield y_2, as x_1 and y_1 are both known at this stage of the process. The second application of (44) yields x_2; the second application of (42) yields y_3; etc.) When the process has been completed, that is, when the approximate value of y_N is calculated, the last member of (44), which has not yet been used, is used as a check to see if these approximate parameters actually define a relative extremum. If the last member of (44) is not satisfied, we select another value for y_1 and repeat the process. If it is satisfied, a relative extremum has been located. The search then continues with the objective of locating any remaining relative extremum on the surface. The entire surface must be searched.

It is a straightforward matter to extend Max's algorithm to the more general case for which the error-weighting function is monotonic. This extended algorithm is identical with the one above, except that the approximate x_i are determined by Eq. 39 and not by its mean-square form Eq. 44.

So much for the case of nonzero $p_x(\xi)$ in the interval $X_\ell < \xi < X_u$. Now consider the case in which $p_x(\xi)$ is zero over one or more subintervals in the interval $X_\ell < \xi < X_u$. Relative extrema will still exist, but in general they will be more numerous and more difficult to locate. The additional difficulty encountered in locating the relative extrema is due to the nature of Eq. 40, that is, to the factor $p_x(x_i)$. No satisfactory algorithm has been obtained, thus far, for determining the relative extrema in this case.

3.3 EXAMPLE

Our objective here is to apply the results just obtained. In particular, we want to determine the optimum two-level quantizer for a signal having the amplitude probability

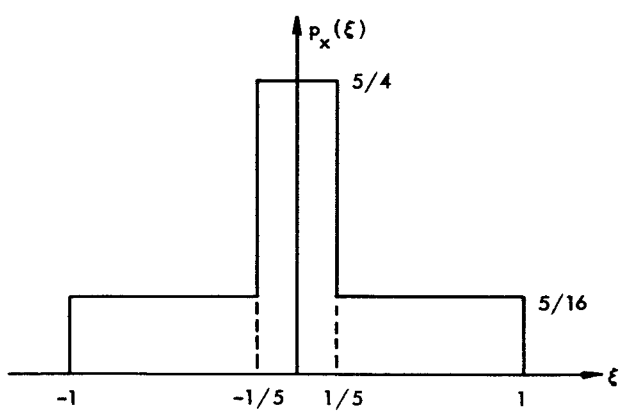

Fig. 6. Amplitude probability density for the example.

density shown in Fig. 6. We choose to optimize in the sense of minimizing the mean-square error.

Successive application of the algorithm above yields three relative extrema:

1. $x_1 = -\frac{1}{5}$

   $y_1 = -\frac{3}{5}$

   $y_2 = \frac{1}{5}$;

2. $x_1 = 0$

   $y_1 = -\frac{7}{20}$

   $y_2 = \frac{7}{20}$;

3. $x_1 = \frac{1}{5}$

$y_1 = -\frac{1}{5}$

$y_2 = \frac{3}{5}$.

If these three relative extrema are examined, we find, first, that the absolute minimum of the surface is specified by the second set of parameters, that is, by

$x_1 = 0$

$y_1 = -\frac{7}{20}$

$y_2 = \frac{7}{20}$.

Second, we find that the two other relative extrema are in reality saddle points.

We shall now consider a modification of the input probability density shown in Fig. 6. The density is modified by decreasing the width of the spike to one-half its former value and increasing its height accordingly. This new density is shown in Fig. 7. If we apply

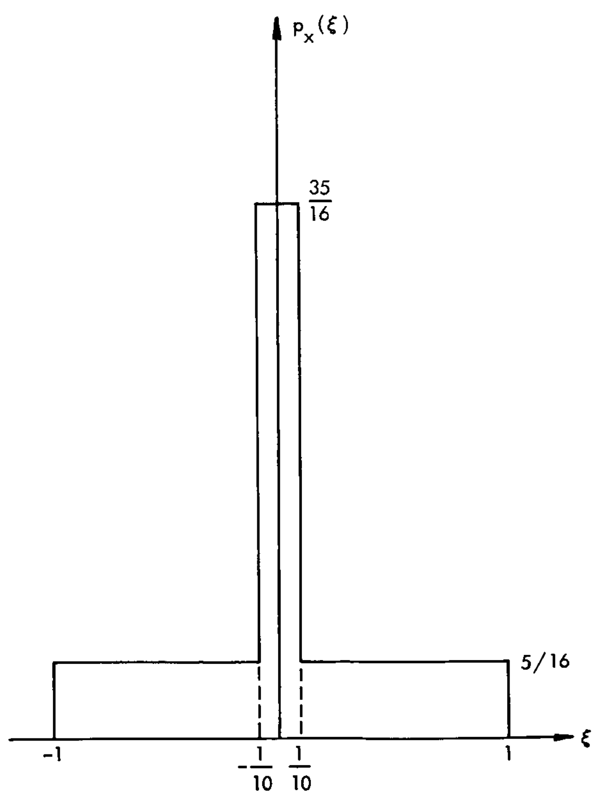

Fig. 7. Modified amplitude probability density.

the algorithm of section 3.2, we again find that the error surface has three relative extrema:

1. $x_1 = -\frac{1}{5}$

   $y_1 = -\frac{3}{5}$

   $y_2 = \frac{1}{5}$

2. $x_1 = 0$

   $y_1 = -\frac{53}{163}$

   $y_2 = \frac{53}{163}$

3. $x_1 = \frac{1}{5}$

   $y_1 = -\frac{1}{5}$

   $y_2 = \frac{3}{5}$.

Investigation indicates that relative extrema one and three are relative minima with identical values for the error. Relative extremum 2 is a saddle point. Its error is greater than that of the two other relative extrema which are therefore the absolute minima of the error surface.

3.4 DISCUSSION

The example just given points out the difficulty that is encountered when we attempt to locate the absolute minima of the error surface by locating all of the relative extrema. Basically, the difficulty is that we do not know how many relative extrema will be located in the search until all points in the region of variation have been considered. Since we expect the number of relative extrema to increase and the search to become more complex as the number of levels in the quantizer is increased, we are forced to conclude that in general this technique is not practical. To be more specific, the algorithm of section 3.2 is of real value only when we can prove that there is only a single relative extremum, a relative minimum, within the region of variation. In the sequel we shall consider three special cases in which the error surface has this single extrema property.

3.5 CONSTRAINED TRANSITION VALUES

Other than the general problem of quantization that we have been considering, there are several quantization schemes of interest. For example, let us assume that the transition values of the quantizer are specified. Such a specification might be required by the quantization equipment, the system containing the quantizer, or some similar constraint.

For such a quantizer, from Eq. 4, we know that the quantization error will be

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{X_i}^{X_{i+1}} d\xi [g(\xi - y_{i+1}) p_x(\xi)]. \tag{45}$$

In (45) the error is a function of the y_i only. We also know that since the y_i are not constrained variables, the absolute minimum of the error surface will be located at a relative extremum.

The relative extrema of this error surface, Eq. 45, are specified by Eq. 39 with fixed transition values, that is, by

$$\int_{X_{k-1}}^{X_k} d\xi \left\{ p_x(\xi) \frac{\partial}{\partial y_k} [g(\xi - y_k)] \right\} = 0, \qquad k = 1, 2, \ldots, N. \tag{46}$$

In writing this equation, we have assumed the error-weighting function g to be continuous. Consider (46) for a moment. Each of the members of this equation contains only one y_k. Therefore, the members may be solved independently to determine the parameters specifying the relative extrema.

In order to satisfy our objectives, we must now determine the error-weighting functions that will yield a single solution (a relative minimum) to Eq. 46. Our method of attack will be to determine a constraint on the error-weighting function which will force every relative extremum to be a relative minimum. Forcing all relative extrema (and saddle points) to be relative minima is sufficient to guarantee that the error surface will have only one relative extremum (a minimum). This single relative minimum will then be the absolute minimum of the surface.

In order to prove that a relative extremum is a relative minimum, we must show that the matrix of second partial derivatives of $\mathscr{E}$ with respect to y_k, k = 1, 2, ..., N, that is,

$$\begin{bmatrix} \frac{\partial^2 \mathscr{E}}{\partial y_1^2} & \frac{\partial^2 \mathscr{E}}{\partial y_1 \partial y_2} & \cdots & \frac{\partial^2 \mathscr{E}}{\partial y_1 \partial y_N} \\ \frac{\partial^2 \mathscr{E}}{\partial y_2 \partial y_1} & \frac{\partial^2 \mathscr{E}}{\partial y_2^2} & \cdots & \frac{\partial^2 \mathscr{E}}{\partial y_2 \partial y_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathscr{E}}{\partial y_N \partial y_1} & \frac{\partial^2 \mathscr{E}}{\partial y_N \partial y_2} & \cdots & \frac{\partial^2 \mathscr{E}}{\partial y_N^2} \end{bmatrix} \tag{47}$$

evaluated at the relative extrema is positive definite. Since the y_k are independent, the off-diagonal terms of (47) are zero; therefore, demonstrating that (47) is positive definite reduces to demonstrating that the N-terms

$$\frac{\partial^2 \mathscr{E}}{\partial {y_k}^2}, \quad k = 1, 2, \dots, N \tag{48}$$

are greater than zero. Referring to Eq. 33 as interpreted for constrained transition values, we have

$$\frac{\partial^2 \mathscr{E}}{\partial {y_k}^2} = \int_{X_{k-1}}^{X_k} d\xi \left\{ p_x(\xi) \frac{\partial^2}{\partial {y_k}^2} [g(\xi - y_k)] \right\}, \quad k = 1, 2, \dots, N. \tag{49}$$

Since $p_x(\xi)$ is positive, a sufficient condition for the members of (49) evaluated at the relative extrema to be positive is for

$$\frac{\partial^2}{\partial {y_k}^2} [g(\xi - y_k)], \quad k = 1, 2, \dots, N \tag{50}$$

to be positive. Functions g for which the members of (50) are greater than zero are called strictly convex functions. Formally, a function g is strictly convex if and only if

$$g[\alpha a + (1-\alpha) b] < \alpha g(a) + (1-\alpha) g(b), \quad \text{for all } b > a \text{ and all } \alpha \text{ such that } 0 < \alpha < 1. \tag{51}$$

Therefore, we can conclude in the case of constrained transition values and strictly convex error-weighting functions that the error surface has a single relative extremum that is a relative minimum. This relative minimum is the absolute minimum of the surface, and is easily located by the method of calculus.

3.6 CONSTRAINED REPRESENTATION VALUES

Another type of quantization that is of interest is the case for which the representation values are specified under the constraint

$$Y_k < Y_{k+1}, \quad k = 1, 2, N-1. \tag{52}$$

By making use of Eq. 4 written for constrained representation values, the error is

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi [g(\xi - Y_{i+1}) p_x(\xi)].$$

If $p_x(\xi)$ is nonzero in the interval $X_\ell < \xi < X_u$, then the relative extrema are solutions of

$$[g(x_i - Y_i) - g(x_i - Y_{i+1})] = 0, \quad i = 1, 2, \dots, N-1 \tag{53}$$

which is Eq. 41 adapted to constrained representation values. Proceeding in a manner analogous to that in section 3.5, we can show that if g is a monotonic error-weighting function (as previously defined), then this error surface will have a single relative extremum that will be a relative minimum and therefore the absolute minimum of the surface. The optimum transition values in this case are specified by Eq. 53.

3.7 CONSTRAINED QUANTIZER-INPUT SIGNALS

In sections 3.5 and 3.6 the structure of the quantizer was constrained and in each case sufficient conditions were found on the error-weighting function so that the error surface will have a single relative extremum, a relative minimum. In a recent paper,[41] P. E. Fleischer has constrained the error-weighting function to be $g(e) = e^2$ and he has derived a sufficient condition on the amplitude probability density of the quantizer-input signal so that, once again, the error surface will have one relative extremum, a relative minimum. Fleischer's sufficient condition is given by the inequality

$$\frac{\partial^2}{\partial \xi^2}\{\ln[p_x(\xi)]\} < 0, \tag{54}$$

where $p_x(\xi)$ is required to be continuous. His method of proof follows the type of arguement that we used in section 3.5, differing only in that he used a row-sum condition[42] to determine the sufficient condition for the matrix of second partials (see Eq. 47) to be positive definite.

The form of Fleischer's condition makes it almost impossible to use for experimentally obtained amplitude probability densities. Referring to Eq. 51, however, we see that (54) is equivalent to requiring that the function

$$\psi(\xi) = -\ln[p_x(\xi)] \tag{55}$$

be strictly convex. Observing that the strictly convex criteria, (51), may be alternatively written

$$e^{-\psi[\alpha a+(1-\alpha)b]} > e^{-[\alpha\psi(a)+(1-\alpha)\psi(b)]} \tag{56}$$

for all $b > a$ and for all α such that $0 < \alpha < 1$, we shall write (55)

$$p_x(\xi) = e^{-\psi(\xi)} \tag{57}$$

and, by direct substitution of (57) in (56), obtain

$$p_x[\alpha a+(1-\alpha)b] > [p_x(a)]^{\alpha}[p_x(b)]^{(1-\alpha)}. \tag{58}$$

If this inequality is satisfied for all $b > a$ and all α such that $0 < \alpha < 1$, then Fleischer's condition is satisfied.

Examination of Eq. 58 indicates several properties of the amplitude probability densities which satisfy this condition. First, if we consider the case in which $p_x(a) = p_x(b)$, we find that (58) can be written

$$p_x[\alpha a+(1-\alpha)b] > p_x(a) = p_x(b) \tag{59}$$

or

$$p_x(\xi) > p_x(a) = p_x(b), \qquad \text{for } a < \xi < b \tag{60}$$

This implies that the $p_x(\xi)$ satisfying this condition must have only one relative extremum and that this relative extremum is a relative maximum. Second, if we consider the case in which $p_x(b) = \beta p_x(a)$, (58) becomes

$$p_x(\xi) > \beta^{(1-\alpha)} p_x(a), \qquad \text{for } a < \xi < b. \tag{61}$$

From a graphical examination of this condition we see that the $p_x(\xi)$ that satisfy Fleischer's condition possess a mild convexity property.

3.8 CONCLUSION

We have shown that under certain conditions on the error-weighting function and the probability density of the quantizer-input signal the optimum quantizer is defined by a relative extremum of the error surface. We then derived equations defining the error surface's relative extrema. In order to apply this technique of determining the optimum quantizer, it is necessary to locate all of the relative extrema of the surface and evaluate the error at each of these points. In most cases, because of the large number of relative extrema expected, this technique is not a practical method of determining the optimum quantizer.

IV. QUANTIZATION OF A SIGNAL CONTAMINATED BY NOISE

We have considered the problem of designing the optimum quantizer for a specified message signal. In many cases, however, the signal of interest is contaminated by noise before reaching the input terminals of the quantizer. We shall now develop an algorithm by which the parameters defining the optimum quantizer for this class of quantizer-input signals can be calculated. Basically, our approach is to treat the quantizer as a non-linear, fixed-form, zero-memory filter. This implies that our objective will be to determine the parameters defining the filter of this class with the minimum error.

We shall consider the relationship that exists between the optimum quantizer for this class of signals and the optimum zero-memory filter when the error-weighting function in each case is

$$g(e) = e^2.$$

We shall also demonstrate how the algorithm developed for quantization can be extended to determine optimum nonlinear, zero-memory filters with other fixed forms.

4.1 FORMULATION OF THE QUANTIZATION PROBLEM

Mathematically, the quantizer-input signal x which consists of the message signal corrupted by noise may be written

$$x = s \oplus n, \tag{63}$$

where s is the message signal, n is the noise, and the symbol $\oplus$ indicates some combination of the two variables, s and n. Two combinations of interest in communication systems are

$$x = s + n$$

and

$$x = s \cdot n.$$

It will be seen from Eq. 65 that any combination $\oplus$ for which a joint probability density of x and s can be defined is an allowable combination.

Proceeding in a manner analogously to the filtering problem, we select as the desired quantizer output the message signal, s. That is, we desire that the quantized signal

$$y = Q(x) = Q(s \oplus n)$$

be an instantaneous replica of the message portion of the quantizer-input signal. In general, we shall demand more than the quantizer can accomplish. There will be an error,

$$e = s - Q(x). \tag{64}$$

We shall take an appropriate mean value of e as a measure of how well the quantizer performs with respect to the demands. This measure of the error is given by

$$\mathscr{E} = \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} d\eta \{g[\eta - Q(\xi)] p_{x,s}(\xi, \eta)\}. \tag{65}$$

$p_{x,s}(\xi, \eta)$ is the joint amplitude probability density of the quantizer-input signal x and the message signal (which is also the desired output signal) s. As in the previous case,

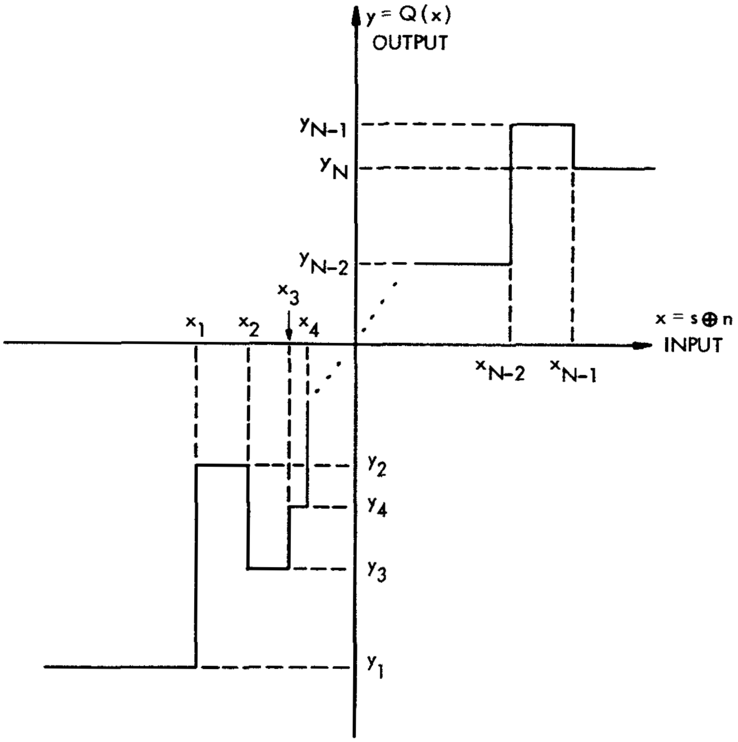

Fig. 8. Input-output relationship of the N-level quantizer.

g is the error-weighting function. The error-weighting function is not required to be either convex or symmetric.

In order to relate the parameters of the quantizer to the error $\mathscr{E}$, we introduce into Eq. 65 the explicit expression for the characteristic of the quantizer (Fig. 8),

$$Q(\xi) = y_k \qquad x_{k-1} \leq \xi < x_k, \qquad k = 1, 2, \ldots, N. \tag{66}$$

(Figure 8 is identical to Fig. 1 except that in Fig. 8 we have made explicit the fact that y_{k-1} is not required to be less than y_k.) Substituting (66) in (65), we have for the quantization error

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \int_{-\infty}^{\infty} d\eta \{g[\eta - y_{i+1}] p_{x,s}(\xi, \eta)\}. \tag{67}$$

Here, x_o is equal to X_ℓ, the greatest lower bound to the quantizer-input signal x; and x_N is equal to X_u, the least upper bound to the quantizer-input signal.

From Eq. 67 it is clear that the error $\mathscr{E}$ is a function of the quantizer parameters $(x_1, x_2, \ldots, x_{N-1}; y_1, y_2, \ldots, y_N)$.

The problem before us then is identical in form to the problem encountered in Section II. That is, the problem is to determine the particular x_i $(i = 1, 2, \ldots, N-1)$ and y_j $(j = 1, 2, \ldots, N)$, the X_i and Y_j that minimize the error $\mathscr{E}$, Eq. 67. Such a minimization is subject to the realizability constraints

$$\left.\begin{aligned} X_\ell = \quad & x_o \leqslant x_1 \\ & x_1 \leqslant x_2 \\ & x_2 \leqslant x_3 \\ & \quad \vdots \\ & x_{N-2} \leqslant x_{N-1} \\ & x_{N-1} \leqslant x_N = X_u \end{aligned}\right\} \tag{68}$$

which are explicit in Fig. 8. This problem is equivalent to that of determining the coordinates of the absolute minimum of the error surface defined by (67) within the region of variation specified by (68).

4.2 QUANTIZATION ALGORITHM

Our objective is to present an algorithm that will permit us to determine the parameters defining the absolute minimum of Eq. 67, subject to the constraints of Eq. 68. Before we consider this algorithm we should compare this error with the error in the case for which the quantizer-input signal is an uncorrupted message signal. The quantization error for an uncorrupted quantizer-input signal is given by Eq. 4. The important thing to observe in comparing these two error expressions, Eqs. 4 and 67, is that they are almost identical in form. Therefore, since the constraints, Eqs. 5 and 68, are identical, we would expect to be able to use the same technique to determine the optimum quantizer in this case as we used for an uncorrupted message signal.

In order to apply the technique that we used earlier, that is, the technique of dynamic programming, it is necessary to define three sets of functionals: error functionals, $\{\epsilon_i(x_i)\}$; transition-value decision functionals, $\{X_i(x)\}$; and representation-value decision functionals, $\{Y_i(x)\}$. The (N) members of each of these three sets of functionals are defined as follows:

$$
\left.\begin{aligned}
\epsilon_1(x_1) &= \min_{\substack{y_1 \\ X_\ell = x_o \leq x_1 \leq X_u}} \left\{ \int_{x_o}^{x_1} d\xi \int_{-\infty}^{\infty} d\eta [g(\eta - y_1)\, p_{x,s}(\xi, \eta)] \right\} \\
\epsilon_2(x_2) &= \min_{\substack{x_1, y_2 \\ X_\ell \leq x_1 \leq x_2 \leq X_u}} \left\{ \epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi \int_{-\infty}^{\infty} d\eta [g(\eta - y_2)\, p_{x,s}(\xi, \eta)] \right\} \\
&\vdots \\
\epsilon_i(x_i) &= \min_{\substack{x_{i-1}, y_i \\ X_\ell \leq x_{i-1} \leq x_i \leq X_u}} \left\{ \epsilon_{i-1}(x_{i-1}) + \int_{x_{i-1}}^{x_i} d\xi \int_{-\infty}^{\infty} d\eta [g(\eta - y_i)\, p_{x,s}(\xi, \eta)] \right\} \\
&\vdots \\
\epsilon_{N-1}(x_{N-1}) &= \min_{\substack{x_{N-2}, y_{N-1} \\ X_\ell \leq x_{N-2} \leq x_{N-1} \leq X_u}} \left\{ \epsilon_{N-2}(x_{N-2}) + \int_{x_{N-2}}^{x_{N-1}} d\xi \int_{-\infty}^{\infty} d\eta [g(\eta - y_{N-1})\, p_{x,s}(\xi, \eta)] \right\} \\
\epsilon_N(x_N) &= \min_{\substack{x_{N-1}, y_N \\ X_\ell \leq x_{N-1} \leq x_N \leq X_u}} \left\{ \epsilon_{N-1}(x_{N-1}) + \int_{x_{N-1}}^{x_N} d\xi \int_{-\infty}^{\infty} d\eta [g(\eta - y_N)\, p_{x,s}(\xi, \eta)] \right\}
\end{aligned}\right\} \tag{69}
$$

$$
\left.\begin{aligned}
&X_1(x) = X_\ell, \text{ a constant;} \\
&X_2(x) = \text{the value of } x_1 \text{ in the coordinate pair } (x_1, y_2) \text{ that minimizes} \\
&\qquad \left\{ \epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi \int_{-\infty}^{\infty} d\eta [g(\eta - y_2)\, p_{x,s}(\xi, \eta)] \right\}, \qquad x_2 = x; \\
&\vdots \\
&X_N(x) = \text{the value of } x_{N-1} \text{ in the coordinate pair } (x_{N-1}, y_N) \text{ that} \\
&\qquad \text{minimizes} \\
&\qquad \left\{ \epsilon_{N-1}(x_{N-1}) + \int_{x_{N-1}}^{x_N} d\xi \int_{-\infty}^{\infty} d\eta [g(\eta - y_N)\, p_{x,s}(\xi, \eta)] \right\}, \qquad x_N = x.
\end{aligned}\right\} \tag{70}
$$

$$
\left.\begin{aligned}
&Y_1(x) = \text{the value of } y_1 \text{ that minimizes}\\
&\qquad \left\{\int_{x_o}^{x_1} d\xi \int_{-\infty}^{\infty} d\eta[g(\eta-y_1)\, p_{x,s}(\xi,\eta)]\right\}, \qquad x_1 = x;\\
&Y_2(x) = \text{the value of } y_2 \text{ in the coordinate pair } (x_1, y_2) \text{ that minimizes}\\
&\qquad \left\{\epsilon_1(x_1) + \int_{x_1}^{x_2} d\xi \int_{-\infty}^{\infty} d\eta[g(\eta-y_2)\, p_{x,s}(\xi,\eta)]\right\}, \qquad x_2 = x;\\
&\qquad \vdots\\
&Y_N(x) = \text{the value of } y_N \text{ in the coordinate pair } (x_{N-1}, y_N) \text{ that minimizes}\\
&\qquad \left\{\epsilon_{N-1}(x_{N-1}) + \int_{x_{N-1}}^{x_N} d\xi \int_{-\infty}^{\infty} d\eta[g(\eta-y_N)\, p_{x,s}(\xi,\eta)]\right\}, \qquad x_N = x.
\end{aligned}\right\} \tag{71}
$$

It follows from Eq. 69 that each of the members of the sets $\{X_i(x)\}$ and $\{Y_i(x)\}$ is defined only over the interval $X_\ell \leq x \leq X_u$.

These three sets of functionals are identical in nature and purpose to those defined in Eqs. 7, 8, and 9. Once they are calculated, the procedure outlined in section 2.2 is used to determine the parameters of the quantizer with minimum error.

4.3 SIMPLIFICATION OF THE ERROR FUNCTIONALS

Consider the k^{th} member of the set of functionals $\{\epsilon_i(x_i)\}$, Eq. 69,

$$
\epsilon_k(x_k) = \min_{\substack{x_{k-1}, y_k \\ X_\ell \leq x_{k-1} \leq x_k \leq X_u}} \left\{\epsilon_{k-1}(x_{k-1}) + \int_{x_{k-1}}^{x_k} d\xi \int_{-\infty}^{\infty} d\eta[g(\eta-y_k)\, p_{x,s}(\xi,\eta)]\right\}. \tag{72}
$$

Referring once again to Eq. 69, we observe that $\epsilon_{k-1}(x_{k-1})$ is not a function of the k^{th} representation value, y_k. By taking this into consideration, (72) may be written alternatively

$$
\epsilon_k(x_k) = \min_{\substack{x_{k-1} \\ X_\ell \leq x_{k-1} \leq x_k \leq X_u}} \left\{\epsilon_{k-1}(x_{k-1}) + \min_{y_k} \int_{x_{k-1}}^{x_k} d\xi \int_{-\infty}^{\infty} d\eta[g(\eta-y_k)\, p_{x,s}(\xi,\eta)]\right\}. \tag{73}
$$

Thus, for specific values of x_k and x_{k-1}, the minimization with respect to y_k can be carried out independently. The specific y_k that minimizes

$$
\int_{x_{k-1}}^{x_k} d\xi \int_{-\infty}^{\infty} d\eta[g(\eta-y_k)\, p_{x,s}(\xi,\eta)] \tag{74}
$$

will be a function of x_k and x_{k-1}. Denoting this value of y_k by y_k^*, we may write (72) as

$$\epsilon_k(x_k) = \min_{\substack{x_{k-1} \\ X_\ell \leq x_{k-1} \leq x_k \leq X_u}} \left\{ \epsilon_{k-1}(x_{k-1}) + \int_{x_{k-1}}^{x_k} d\xi \int_{-\infty}^{\infty} d\eta [g(\eta - y_k^*)\, p_{x,s})\xi, \eta)] \right\}. \tag{75}$$

Comparing Eqs. 72 and 75, we see that the effect of separating the two minimizations is to reduce the formal search, which is necessary to obtain the error functionals, from a two-dimensional to a one-dimensional search.

A very pertinent question now concerns determination of the value of y_k^*. Recall that when the quantization problem was originally stated we noted that the y_k were not constrained variables. Thus, the absolute minimum of (74) with respect to y_k must be a relative extremum. If g is a continuous function, the relative extrema and therefore y_k^* must be solutions of the equation

$$\int_{x_{k-1}}^{x_k} d\xi \int_{-\infty}^{\infty} d\eta \left\{ p_{x,s}(\xi, \eta) \frac{\partial}{\partial y_k} [g(\eta - y_k)] \right\} = 0. \tag{76}$$

It can be shown that if g, as well as being continuous, is a convex error-weighting function and if $p_{x,s}(\xi,\eta)$ is nonzero over some subinterval in the interval defined by $x_{k-1} \leq \xi < x_k$ and $-\infty \leq \eta \leq \infty$, then Eq. 76 has only one solution. This solution is a relative minimum and is therefore the value of y_k which we have called y_k^*. Thus it should be evident that in this case the labor involved in obtaining the error functionals is greatly reduced.

For noncontinuous error-weighting functions, y_k^* is determined by a direct search along the set of possible values for y_k. In this case there is no reduction in the labor required to determine the error-weighting functionals.

4.4 A SECOND VIEW OF THE QUANTIZATION PROBLEM

Our objective now is to derive an alternative expression for the quantization error when the quantizer-input signal is a message signal corrupted by noise. In section 4.1 we found the measure of the quantization error (Eq. 67). Now suppose we let the quantization error be denoted by the random variable λ. That is,

$$\begin{aligned} \lambda &= \eta - Q(\xi) \\ &= \eta - y_i \qquad x_{i-1} \leq \xi < x_i, \qquad i = 1, 2, \ldots, N. \end{aligned} \tag{77}$$

Solving (77) for η and substituting this in (67), we obtain

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \int_{-\infty}^{\infty} d\lambda [g(\lambda)\, p_{x,s}(\xi, \lambda + y_{i+1})]. \tag{78}$$

If the terms of (78) are rearranged by interchanging the order of integration with respect to λ with the summation on i and the integration with respect to ξ, we have

$$\mathscr{E} = \int_{-\infty}^{\infty} d\lambda\ g(\lambda) \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi\ p_{x,s}(\xi, \lambda+y_{i+1}). \tag{79}$$

From (79) it is apparent that the term

$$\sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi\ p_{x,s}(\xi, \lambda+y_{i+1}) \tag{80}$$

is a function of λ and the quantizer parameters. Let us now interpret this term. Consider the k^{th} term of the sum, Eq. 80. An examination of this k^{th} term indicates that it represents the contribution to the amplitude probability density of the error signal by that portion of the quantizer-input signal having amplitudes in the range $x_{k-1} \leqslant \xi < x_k$. Thus, since the amplitude regions $x_{k-1} \leqslant \xi < x_k$, $k = 1, 2, \ldots, N$, are mutually exclusive, (80) is the amplitude probability density of the error signal. That is,

$$p_e(\lambda) = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi\ [p_{x,s}(\xi, \lambda+y_{i+1})]. \tag{81}$$

Substituting Eq. 81 in Eq. 79, we have

$$\mathscr{E} = \int_{-\infty}^{\infty} d\lambda [g(\lambda)\ p_e(\lambda)]. \tag{82}$$

Equation 82 indicates, as did Eq. 32 for the uncorrupted signal case, that the problem of designing the optimum quantizer is equivalent to the shaping of the amplitude probability density of the error signal so that the property indicated by the error-weighting function g is minimized. This operation of shaping the error density is constrained by the number of levels (N) and by the joint amplitude probability density of the quantizer-input signal x and the message signal s.

4.5 THE NATURE OF THE ABSOLUTE MINIMUM

For the case of an uncorrupted quantizer-input signal we were able to prove (section 3.1) for rather general conditions that the absolute minimum of the error surface will be located at one of the error surface's relative minima. A similar result has not been obtained for the case in which the quantizer-input signal is a corrupted message signal. The primary reason that such a result cannot be obtained in this case lies in the additional complexity of the problem, because of the corrupting effect of the noise. In fact, it is not feasible, because of the nature of the equations specifying the relative extrema, to define an algorithm that permits us to locate these relative extrema of the error surface.

There is, however, one case of interest where the error surface has only one relative extremum, a relative minimum. This relative minimum, which is also the absolute minimum, can be found by using the methods of calculus. We shall consider this special case.

4.6 CONSTRAINED TRANSITION VALUES

When the quantizer-input signal is a message contaminated by noise the quantization error for constrained transition values is

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{X_i}^{X_{i+1}} d\xi \int_{-\infty}^{\infty} d\eta \{g[\eta - y_{i+1}] p_{x,s}(\xi, \eta)\}. \tag{83}$$

(Equation 83 is identical to Eq. 67 with fixed transition values.) The error in (83) is a function of the y_i only.

Making use of Eq. 38 in connection with Eq. 83, we find that the representation values are specified by solutions of the set of simultaneous equations

$$\int_{X_{k-1}}^{X_k} d\xi \int_{-\infty}^{\infty} d\eta \left\{ \left[\frac{\partial}{\partial y_k} g(\eta - y_k) \right] p_{x,s}(\xi, \eta) \right\} = 0, \qquad k = 1, 2, \ldots, N. \tag{84}$$

In writing (84), we have assumed the error-weighting function g to be continuous. Since each of the members of (84) contains but a single y_k, they may be solved independently to determine the parameters that specify the relative extrema.

Proceeding in a manner identical to that of section 3.5, we find that Eq. 83 will have a unique solution if the error-weighting function is strictly convex. This relative minimum is the absolute minimum of the error surface and is easily located by using the methods of calculus.

4.7 SPECIAL RESULTS FOR $g(e) = e^2$

At this point in our discussion we would like to turn our attention to a problem which, at first, may seem completely unrelated to the problem that is being considered here.

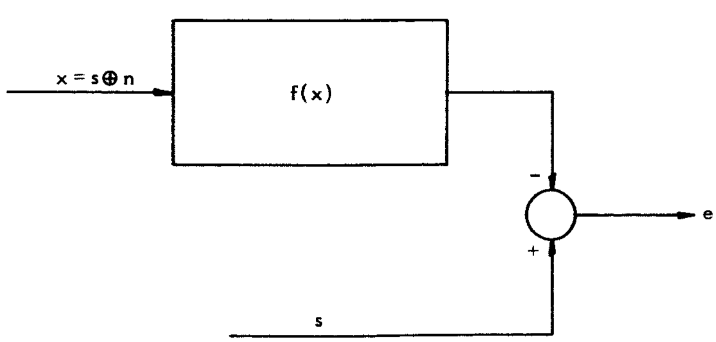

Fig. 9.
Concerning the optimum nonlinear zero-memory filter.

Consider the block diagram of Fig. 9. First of all, we want to determine the specific function $f(\xi)$ that minimizes the mean-square value of the error,

$$\mathscr{E} = \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} d\eta \left\{ [\eta - f(\xi)]^2 \, p_{x,s}(\xi, \eta) \right\}. \tag{85}$$

The specific f(ξ) denoted by F(ξ) which minimizes (85) is given by the conditional mean,

$$F(\xi) = \int_{-\infty}^{\infty} d\eta \left\{ \eta p_{s|x}(\eta|\xi) \right\}. \tag{86}$$

(For a derivation of this result see, for example, Cramér.[43]) By substitution of Eq. 86 in Eq. 85, we find that the minimum value for the error is

$$\mathscr{E}_{\min} = \int_{-\infty}^{\infty} d\eta [\eta^2 p_s(\eta)] - \int_{-\infty}^{\infty} d\xi [F^2(\xi) \, p_x(\xi)]. \tag{87}$$

Assume that this optimum filter F(ξ) is approximated in the minimum mean-square-error sense by an N-step approximation. We want to determine the relationship between the error resulting from this step approximation to F(ξ) and the error resulting from minimum mean-square-error quantization with N-levels.

In order to obtain the desired relation, we first compare the error between step approximation with parameters x_i and y_j and quantization with the same parameters. (These parameters are not required to be either the optimum step parameters or the optimum quantizer parameters.) The error resulting from this step approximation is

$$\mathscr{E}_F = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} [F(\xi) - y_{i+1}]^2 \, p_x(\xi) \, d\xi. \tag{88}$$

Substituting the optimum filter characteristic (86) in (88) and expanding, we obtain

$$\begin{aligned}
\mathscr{E}_F = & \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \left\{ \int_{-\infty}^{\infty} d\eta [\eta p_{s|x}(\eta|\xi)] \right\}^2 p_x(\xi) \, d\xi \\
& - 2 \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \left\{ y_{i+1} \int_{-\infty}^{\infty} d\eta [\eta p_{s|x}(\eta|\xi)] \right\} p_x(\xi) \, d\xi \\
& + \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} y_{i+1}^2 p_x(\xi) \, d\xi.
\end{aligned} \tag{89}$$

In like manner the quantization error for the same parameters x_i and y_j is

$$\mathscr{E}_Q = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \int_{-\infty}^{\infty} d\eta \left[(\eta-y_{i+1})^2 \, p_{x,s}(\xi,\eta)\right] \tag{90}$$

$$= \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \int_{-\infty}^{\infty} d\eta \left[\eta^2 p_{x,s}(\xi,\eta)\right] - 2 \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \int_{-\infty}^{\infty} d\eta [y_{i+1}\eta p_{x,s}(\xi,\eta)]$$

$$+ \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \int_{-\infty}^{\infty} d\eta \left[y_{i+1}^2 p_{x,s}(\xi,\eta)\right]. \tag{91}$$

Comparing Eqs. 87, 89, and 91, we see that

$$\mathscr{E}_Q = \mathscr{E}_{min} + \mathscr{E}_F. \tag{92}$$

Thus, since $\mathscr{E}_{min}$ is a constant, the parameters x_i and y_j that minimize $\mathscr{E}_Q$ also minimize $\mathscr{E}_F$. Therefore, the optimum step approximation to $F(\xi)$ is identical to the minimum mean-square-error quantizer. (A result somewhat similar to this has been obtained by D. A. Chesler.[44])

4.8 OTHER FIXED-FORM, NONLINEAR, ZERO-MEMORY FILTERS

We have considered the problem of designing optimum quantizers when the quantizer-input signal is a message signal corrupted by noise. In this section we shall demonstrate that the quantization algorithm can be extended to design other types of fixed-form, nonlinear, zero-memory filters.

Let us consider the problem of designing the optimum piecewise-linear filter, that is, a filter having an input-output characteristic specified by

$$y = H(\xi) = m_k\xi + b_k \qquad x_{k-1} \leq \xi < x_k \qquad k = 1, 2, \ldots, N. \tag{93}$$

These linear segments are not required to be connected. The transfer characteristic of this filter is pictured in Fig. 10.

Proceeding in a manner identical to that of sections 4.1 and 4.2, we obtain for the filter error

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \int_{-\infty}^{\infty} d\eta \{g[\eta-(m_{i+1}\xi+b_{i+1})]\, p_{x,s}(\xi,\eta)\}. \tag{94}$$

In order to specify the optimum filter of this form, we must minimize (94) with respect to the parameters x_i, m_j, and b_k, subject to the constraints (68). The m_j and b_k are unconstrained variables.

An algorithm which will permit us to determine the optimum $H(\xi)$ is easily formulated

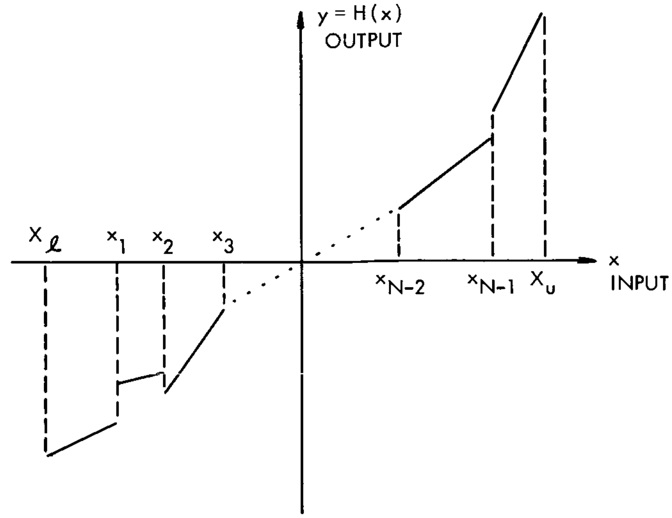

Fig. 10. Input-output characteristic of a piecewise-linear filter.

by using the techniques previously employed in the quantization case. Note, however, that in this case there will be three sets of decision functionals instead of two as in the quantization case.

Suppose that we desire that our piecewise-linear filter be continuous. That is, we want the filter parameters to satisfy

$$m_k x_k + b_k = m_{k+1} x_k + b_{k+1} \qquad k = 1, 2, \ldots, N-1. \tag{95}$$

In order to determine this optimum filter Eq. 95 must be minimized subject to the constraints expressed by Eqs. 68 and 95. In general, a set of additional constraints such as Eq. 95, will not complicate the application of the algorithm. In this particular problem the additional constraints will actually simplify the problem, since they establish a relationship between m_k and b_k at each level, thereby reducing the dimensionality of the problem.

A careful examination of the material presented in section 4.7 for the error-weighting function $g(e) = e^2$, indicates that the result obtained there for the quantizer is also valid for the piecewise-linear filters just discussed.

It should be clear from our discussion that the type of algorithm obtained for the quantizer and the piecewise-linear filters can also be obtained for any piecewise-polynomial filter,

$$H_p(\xi) = a_{ok} + a_{1k}\xi + a_{2k}\xi^2 + \ldots + a_{pk}\xi^p \qquad x_{k-1} \leq \xi < x_k \qquad k = 1, 2, \ldots, N. \tag{96}$$

In addition to the realizability constraints, Eq. 68, which must be applied, up to (p-1) constraints concerning continuity in value, continuity in slope, etc. may be included in the formulation of the algorithm. It can be shown that the result of section 4.7, as well as being valid for the quantizer and the piecewise-linear filter, is also valid for the general polynomial filter, Eq. 96.

V. A COMPUTER STUDY

We shall demonstrate the application of the quantization algorithm that was developed in Section II. In particular, we shall use a version of this algorithm, adapted for computer use, to determine optimum quantizers for a specific quantizer-input signal. Several error-weighting functions will be considered. The computer version of the quantization algorithm is discussed in detail in Appendix C.

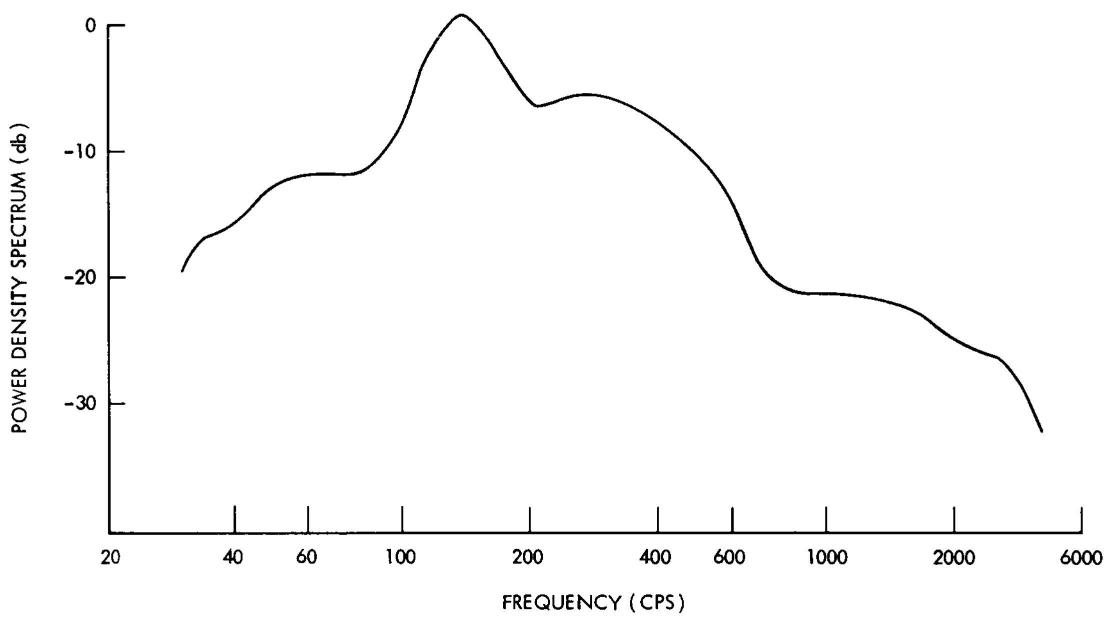

Fig. 11. Power density spectrum of the quantizer-input signal.

The specific signal that is used in this study consists of the two spoken sentences, "Joe took father's shoe bench out. She was waiting at my lawn."[45] These two sentences contain most of the important phonemes and have a frequency spectrum (see Fig. 11) that is roughly typical of conversational speech. This sample was collected by Ryan[46] who used it in a preliminary study of the optimum quantization of speech. The central portion of the amplitude probability density of this sample is shown in Fig. 12. It should be noted that this short sample (the two-sentence sample was approximately four seconds in length) amplitude probability density is almost identical to published data[47] on long-sample amplitude densities.

After the selection of an input signal for quantization it is necessary to select a specific error-weighting function. For purposes of this example we shall consider the following specific situations:

1. Transition values constrained to be uniformly spaced and representation values chosen to minimize the mean-square error;

2. Representation and transition values chosen to minimize the mean-square value of the error;

3. Representation and transition values chosen to minimize the mean-absolute value of the error;

4. Representation and transition values chosen to minimize the mean-square value of the percentage error; that is, the Y_j and X_i are chosen to minimize

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \left[\left(\frac{\xi - y_{i+1}}{y_{i+1}}\right)^2 p_x(\xi)\right]. \tag{97}$$

(The last quantization scheme illustrates the adaptability of the quantization algorithm to other definitions of the error, e.)

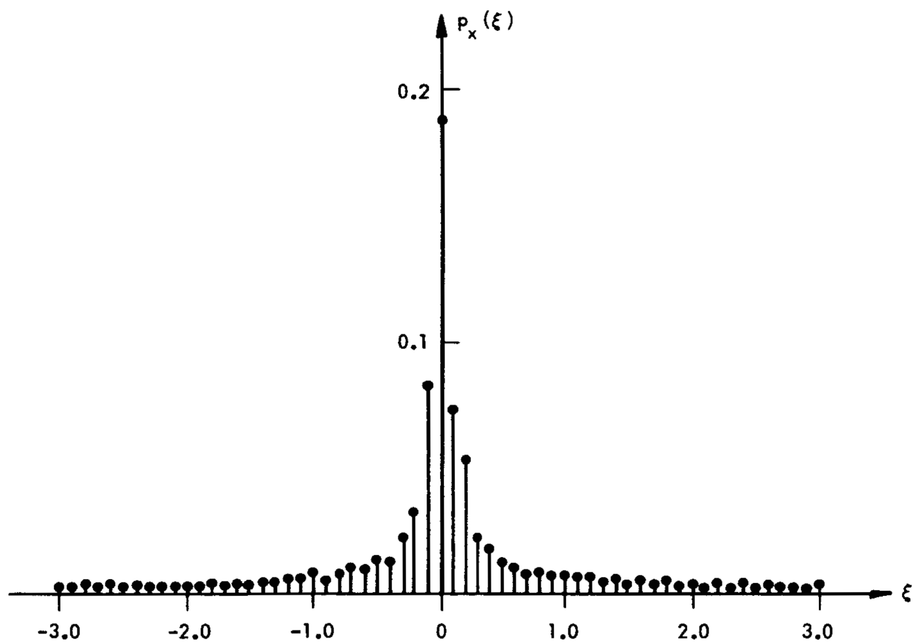

Fig. 12. Central portion ($-3 \leq \xi \leq 3$) of the amplitude probability density of the speech sample. (The original signal x(t) is so bounded that $-12.8 \leq x(t) \leq 12.8$ for all t.)

The quantization algorithm has been programmed on the IBM 7094 digital computer for these four quantization schemes. Typical of the results that are obtained through the application of these computer programs are those obtained for eight-level quantization. Table 1 presents the parameters that define these optimum quantizers, together with the parameters that define the eight-level uniform quantizer and the eight-level logarithmatic quantizer ($\mu = 100$). (See Smith[9] for a definition of logarithmetic quantization.) A comparison of the columns of Table 1 or Fig. 13 illustrates that the optimum

Table 1. Eight-level quantizer parameters.

	Uniform quantization	Logarithmetic quantization, $\mu = 100$	Constrained transition value, minimum mean-square representation value quantization	Minimum mean-square error quantization	Minimum mean-absolute error quantization	Minimum mean-square percentage error quantization
y_1	-11.20	-7.13	-10.91	-10.01	-7.60	-9.45
x_1	-9.60	-3.95	-9.60	-8.05	-5.85	-6.95
y_2	-8.00	-2.16	-7.61	-6.10	-4.10	-5.00
x_2	-6.40	-1.16	-6.40	-4.75	-3.05	-3.35
y_3	-4.80	-0.59	-4.45	-3.38	-2.00	-2.17
x_3	-3.20	-0.28	-3.20	-2.35	-1.35	-1.05
y_4	-1.60	-0.10	-0.51	-1.32	-0.70	-0.57
x_4	0.00	0.00	0.00	-0.65	-0.35	-0.05
y_5	1.60	0.10	0.89	0.04	0.00	0.34
x_5	3.20	0.28	3.20	0.85	0.45	0.65
y_6	4.80	0.59	4.38	1.64	0.90	1.55
x_6	6.40	1.16	6.40	2.65	1.65	2.35
y_7	8.00	2.16	7.60	3.68	2.50	3.53
x_7	9.60	3.95	9.60	5.25	3.65	4.95
y_8	11.20	7.13	10.67	6.79	4.80	6.83

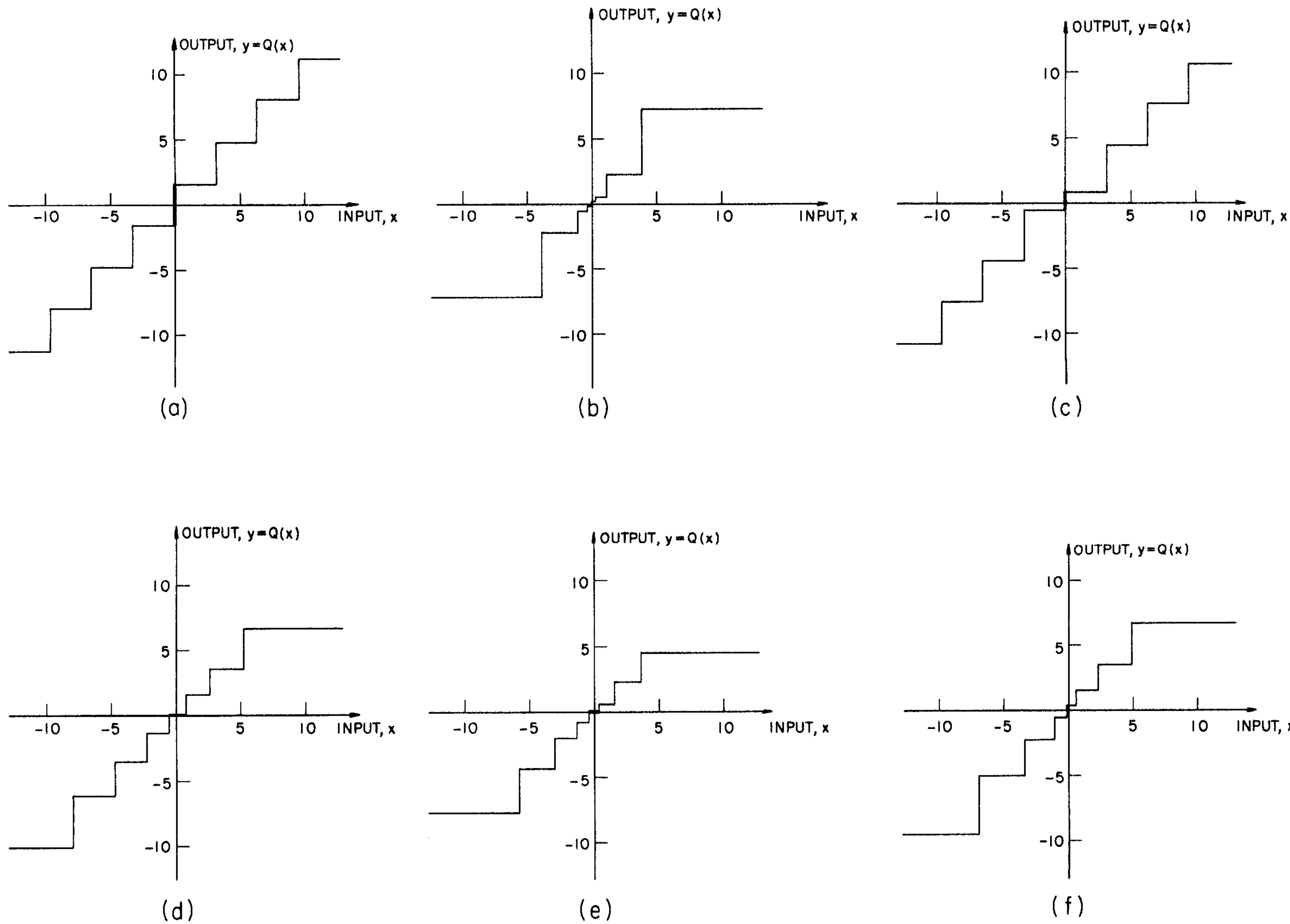

Fig. 13. (a) Eight-level uniform quantizer. (b) Eight-level logarithmetic quantizer, $\mu = 100$. (c) Eight-level quantizer with constrained transition values and representation values chosen to minimize the mean-square error. (d) Eight-level minimum mean-square-error quantizer. (e) Eight-level minimum mean-absolute-error quantizer. (f) Eight-level minimum mean-square-percentage-error quantizer.

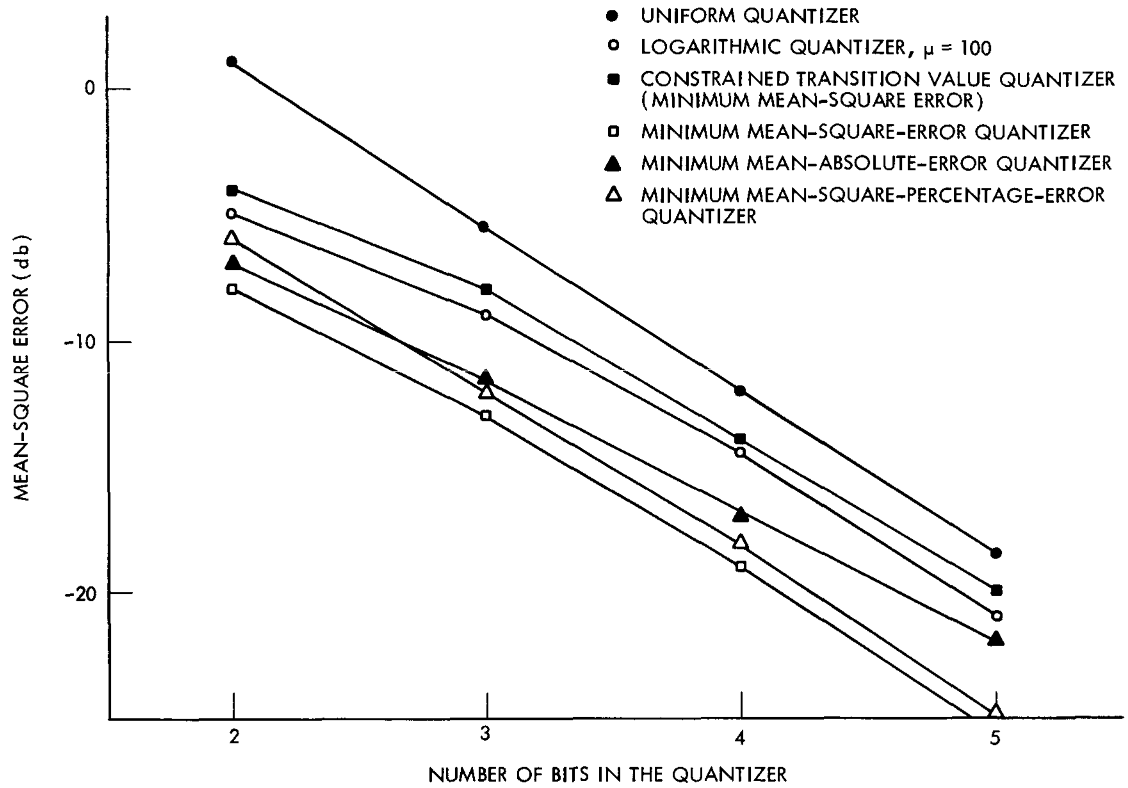

Fig. 14. Mean-square error versus the number of levels (in bits). (Zero db is the mean-square value of the input signal.)

quantizers tend to place more levels in the regions of high probability. The exact nature of the concentration of the levels depends on the error-weighting function, the amplitude probability density of the quantizer-input signal, and the quantization scheme. Figure 14 compares the mean-square value of the quantization error for these six types of quantizers as a function of the number of quantization levels.

In section 2.4 it was shown that the process of determining the optimum quantizer is equivalent to shaping the amplitude probability density of the error signal in such a manner that some property of this density specified by the error-weighting function g is minimized. This being the case, we expect these error probability densities to present a good picture of how the optimum quantizers achieve their reduction in error. Figure 15 pictures the error amplitude probability densities of the six eight-level quantizers pictured in Fig. 13.

In any optimum signal-processing system it is of interest to consider how the error varies when a signal other than the "designed for" signal is applied to the system input. Figure 16 is a plot of the normalized quantization error for a number of quantizers (eight-level quantization) versus an amplitude scaling factor that was used to modify the original quantizer-input signal. In each case the value of the normalized quantization error that is plotted is the actual quantization error divided by the mean-square value of the quantizer-input signal which yielded that error. Each of the curves is normalized

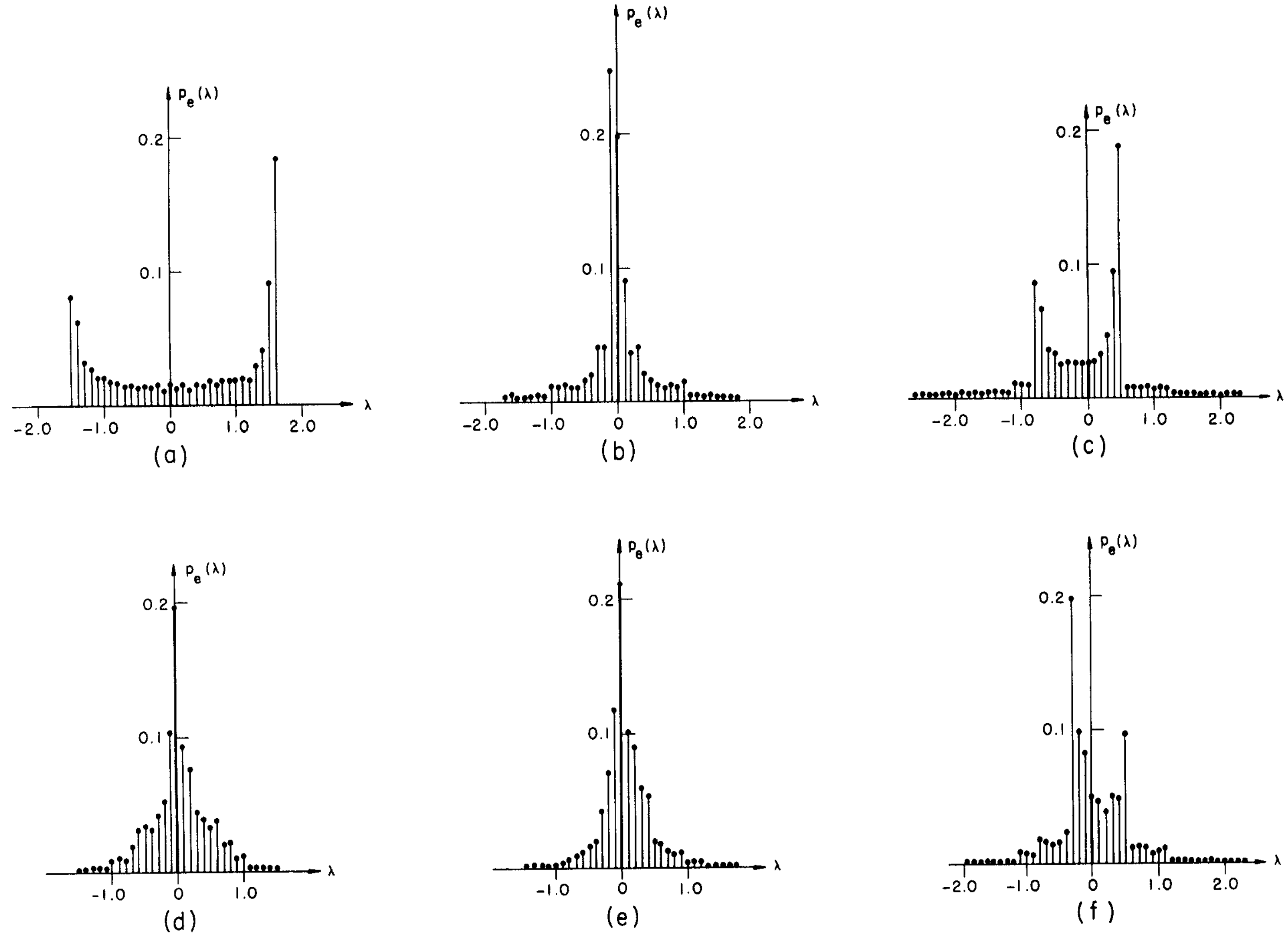

Fig. 15. (a) Amplitude probability density of the error signal, eight-level uniform quantization. (b) Amplitude probability density of the error signal, eight-level logarithmetic quantization, $\mu = 100$. (c) Amplitude probability density of the error signal, eight-level quantization with constrained transition values and representation values chosen to minimize the mean-square value of the error. (d) Amplitude probability density of the error signal, eight-level minimum mean-square error quantization. (e) Amplitude probability density of the error signal, eight-level minimum mean-absolute-error quantization. (f) Amplitude probability density of the error signal, eight-level minimum mean-square-percentage-error quantization.

to zero db independently. Observation of this figure indicates that we can expect more variation in the value of the normalized mean-square quantization error in the case of uniform quantization than in any of the optimum quantization cases studied. And, in

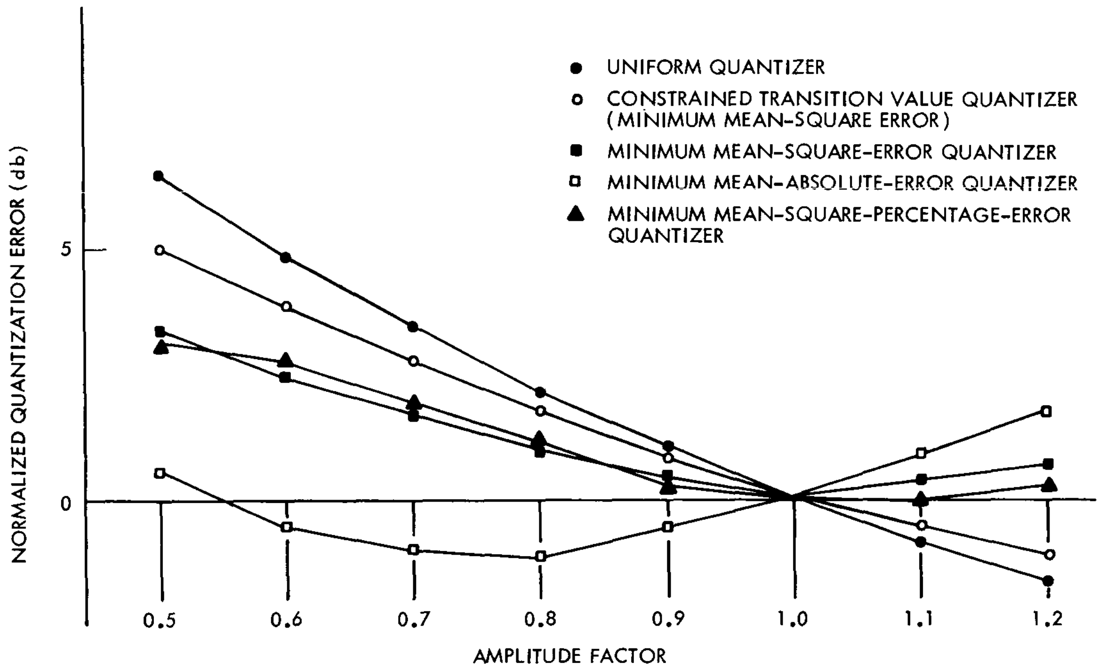

Fig. 16. Normalized quantization error versus amplitude scaling factor, eight-level quantization.

particular, the three cases in which we are at liberty to optimize over both transition and representation values show the least variation.

Results concerning the power density spectrum of the quantizer-output signal and the associated error signal for the case in which the quantizer-input signal is speech have been presented in a thesis by G. Crimi.[48]

VI. CRITIQUE AND EXTENSIONS

In this report we have developed an approach to the problem of designing optimum quantizers for both signals uncontaminated by noise and signals contaminated by noise. In both instances the approach was based on the philosophy that the quantizer is a nonlinear, zero-memory filter of a fixed basic form. The algorithms developed here provide a method of obtaining the parameters that define optimum quantizer, given the amplitude probability density of the signal (joint amplitude probability density of the quantizer-input signal and the noise for the contaminated signal case) and a suitably chosen error-weighting function. Several observations concerning the nature of these algorithms may be made.

First, because of the nature of the minimization problem, which is due to the possibility of boundary solutions, it is necessary to obtain an algorithm that searches for the absolute minimum within the region of variation, rather than one that searches for relative minima. The algorithm thus obtained is applicable for convex and nonconvex error-weighting functions and for discrete and continuous amplitude probability densities. Second, one observes from the formulation of the error functionals that after an initial set of computations the computation time required to calculate the parameters specifying the N-level optimum quantizer is directly proportional to (N−1).

The work presented in this report also suggests two possible areas of future research. We have shown that the quantization algorithm can be extended to other types of fixed-form nonlinear, zero-memory filtering. One interesting area for further study is the possible extension of this algorithm to the design of optimum nonlinear systems with finite memory. The second suggestion for further study comes from the alternative formulation of the quantization error. This alternative formulation suggests that the problem[49,50] of simultaneously designing a discrete signal which will be corrupted by noise and a nonlinear, zero-memory filter in the form of a quantizer can be approached by using an algorithm similar to the quantization algorithm.

APPENDIX A

Dynamic Programming

We shall now investigate on a basic level the technique called "dynamic programming."[37,38] Our purpose is to obtain a working definition of this technique and to apply this definition to a very simple allocation problem.

A. 1 INTRODUCTION

Basically, dynamic programming is a technique for solving a large class of problems which either are or can be transformed into multistage decision processes. (Multistage decision processes are sometimes called "multistage allocation processes.") A problem is considered a multistage decision process if it can be formulated in such a manner that the parameter values that define the solution of the problem can be determined one at a time. (Many of the problems in the calculus of variations can be formulated as multistage decision processes.[51]) The decisions in determining this solution are made according to some well-defined criterion. This criterion is usually expressed as a maximization or minimization of a function of the parameters defining the process. In general these process-defining parameters are subject to some set of constraints. The set of values of these parameters which satisfies all of the constraints is known as the region of variation.

In applying the technique of dynamic programming to a particular problem the primary objective is to imbed the problem of interest in a family of similar problems in such a manner that a complicated process is decomposed into a number of relatively simple processes. In order to investigate this technique we shall consider a simple allocation problem.

A. 2 SIMPLE ALLOCATION PROBLEM

In our allocation problem we assume that a sum of money X is available to be invested either in part or in full in (N) activities A_i, $i = 1, 2, \ldots, N$. If x_i is the

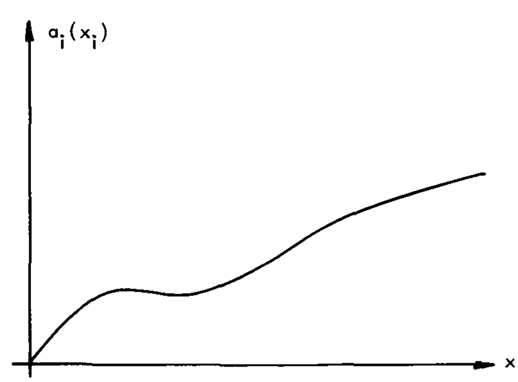

Fig. A-1. A possible return function.

allocation to the i^{th} activity the return from this activity will be given by $a_i(x_i)$. Figure A-1 pictures a possible return function.

Our objective in this allocation problem is to maximize the return from our investment. In order to proceed in the mathematical formulation of the problem we make three assumptions concerning the investment procedure.

(i) The returns from each of the investments can be measured in a common unit.

(ii) The return from any activity is independent of the allocations to the other activities.

(iii) $a_i(0) = 0$ for all i.

Applying these assumptions to the problem, we have for R, the total return,

$$R(x_1, x_2, \ldots, x_N) = \sum_{i=1}^{N} a_i(x_i). \tag{A. 1}$$

This total return from our investment is subject to the constraint

$$\sum_{i=1}^{N} x_i = X, \tag{A. 2}$$

where X is the amount to be invested and the set of constraints

$$x_i \geqslant 0 \qquad i= 1, 2, \ldots, N. \tag{A. 3}$$

Equation A. 2 limits the total investment to the amount of resources available. Equation A. 3 limits each of the allocations to a positive quantity. This set of constraints is necessary, since the concept of negative investments is not defined.

Specifically, we want to determine the maximum value of the return, A. 2, and the values of the x_i yielding this maximum return for any investment x, $0 \leqslant x \leqslant X$.

At this point in our discussion a logical question to ask is, Why not use the methods of calculus to determine the solution? To this, we might reply in the following way. When we apply calculus to the allocation problem (or to problems of a similar structture) we find one problem that calculus cannot surmount: The absolute maxima (or minima when allocations are made on the basis of a functional minimization) will frequently be at a point on the boundary of the region of variation. Generally, this point will not be a relative extremum. If the boundary solution is not a relative extremum, the slope at the point will not be zero and cannot be found by using the methods of calculus.

Since we do not know a priori whether the solution will be at a relative extremum or on the boundary, a solution obtained by using the methods of calculus may be incorrect. To insure a correct solution to the allocation problem, we must employ a technique that searches for the absolute maximum of the surface within the region of variation. Dynamic programming is such a technique.

A.3 A GRAPHICAL SOLUTION TO THE ALLOCATION PROBLEM

Before we formulate the allocation algorithm in equation form we want to consider it as a graphical search technique. We shall begin by considering the first two activities, A_1 and A_2, which have return functions $a_1(x_1)$ and $a_2(x_2)$, respectively. Making use of the results of the preceeding section, we find that the total return from these two activities is

$$R(x_1, x_2) = a_1(x_1) + a_2(x_2) \tag{A.4}$$

and is subject to the constraints

$$x_1 + x_2 = x \tag{A.5}$$

and

$$\left.\begin{array}{l} x_1 \geqslant 0 \\ x_2 \geqslant 0 \end{array}\right\} . \tag{A.6}$$

We want to determine for each x in the range $0 \leqslant x \leqslant X$ the maximum return and the values of x_1 and x_2 which yield this maximum return.

One method of determining the maximum value of the return for a specific x, say $x = \alpha$, is to search the return surface along that portion of the line

$$x_1 + x_2 = \alpha$$

contained in the region of variation. This search path is indicated in Fig. A-2, where $R(x_1, x_2)$ is plotted along an axis directed out of the page. By examining each of the

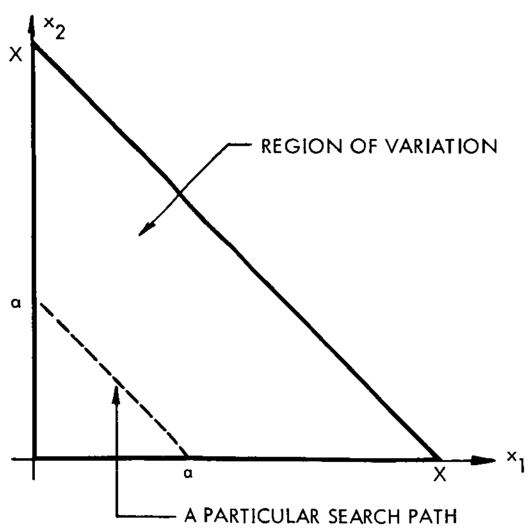

Fig. A-2. Typical search path in the two-activity allocation problem.

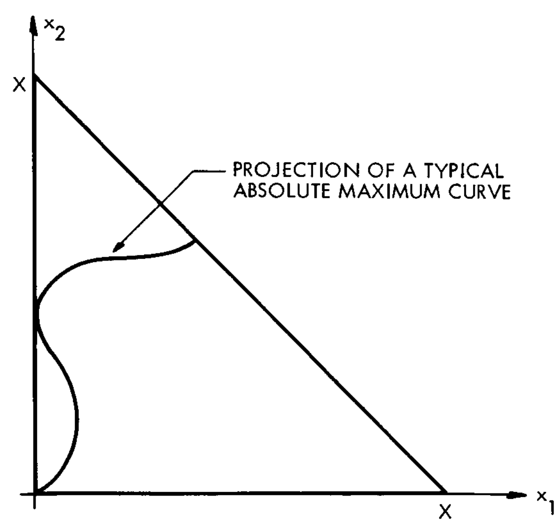

Fig. A-3. Typical projection of the absolute maximum onto the (x_1, x_2)-plane for the two-activity problem.

points on the search path the absolute maximum of the return surface for this specific allocation is easily determined. If this process is repeated for all x in the range $0 \leqslant x \leqslant X$, a curve such as that shown in Fig. A-3 is obtained which indicates the x_1 and x_2 that maximize the return for a specific x.

Now, consider the problem of maximizing the return from the first three activities, A_1, A_2, and A_3. The return from the investment of x_1 in activity A_1, x_2 in A_2, and x_3 in A_3 is

$$R(x_1, x_2, x_3) = a_1(x_1) + a_2(x_2) + a_3(x_3) \tag{A. 7}$$

and is subject to the constraints

$$x_1 + x_2 + x_3 = x \tag{A. 8}$$

and

$$\left.\begin{aligned} x_1 &\geqslant 0 \\ x_2 &\geqslant 0 \\ x_3 &\geqslant 0 \end{aligned}\right\} . \tag{A. 9}$$

We want to determine for each x in the range $0 \leqslant x \leqslant X$ the maximum return and the values of x_1, x_2, and x_3 which yield this maximum return.

Upon first inspection it appears that in this three-activity case we must employ some type of three-dimensional search technique in order to obtain the desired solution. Consider such a three-dimensional search. It will be a search along the plane, (A. 8), for a specific value of x, say $x = \alpha$. Such a search might be conducted by assuming a value for x_3 consistent with the constraints, and then determining the optimum allocation of the remaining resources between x_1 and x_2. This operation would be repeated for each x_3 satisfying the constraints, that is, for x_3 satisfying the inequality

$$0 \leqslant x_3 \leqslant \alpha.$$

From an examination of the results of these calculations we obtain the absolute maximum for this particular value of x, $x = \alpha$.

We should observe, however, that once a value for x_3 is selected, there remains the amount $x - x_3$ (or to use the value of x used previously, $\alpha - x_3$) to be allocated to the activities A_1 and A_2. This is the same problem, however, as the two-activity allocation problem just considered. The basic difference is that now instead of investing an amount x in the two activities we invest $x - x_3$. Since x_3 is positive or zero we know that

$$x - x_3 \leqslant x \leqslant X$$

and since the optimum two-activity allocation problem has been solved for all investments in the range $0 \leqslant x \leqslant X$, we can use the results of this solution without further search to determine the maximum return for the first two activities in the three-activity problem. It is possible to make use of this solution because the two-activity return surface $R(x_1, x_2)$ is a subsurface of the three-activity return surface $R(x_1, x_2, x_3)$. That is,

$$R(x_1, x_2) = R(x_1, x_2, 0) \tag{A.10}$$

since $a_i(o) = 0$ for $i = 1, 2, \ldots, N$. Thus, by making use of our prior knowledge, we are able to reduce the three-dimensional search to a two-dimensional search. The resulting

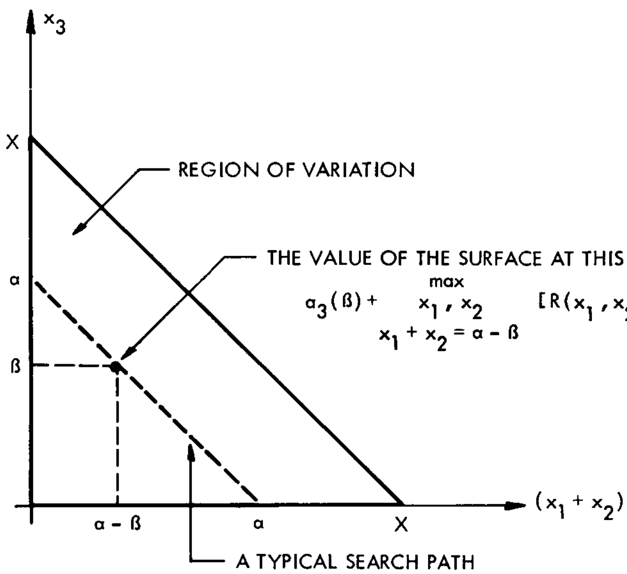

Fig. A-4. Typical search path in the three-activity allocation problem.

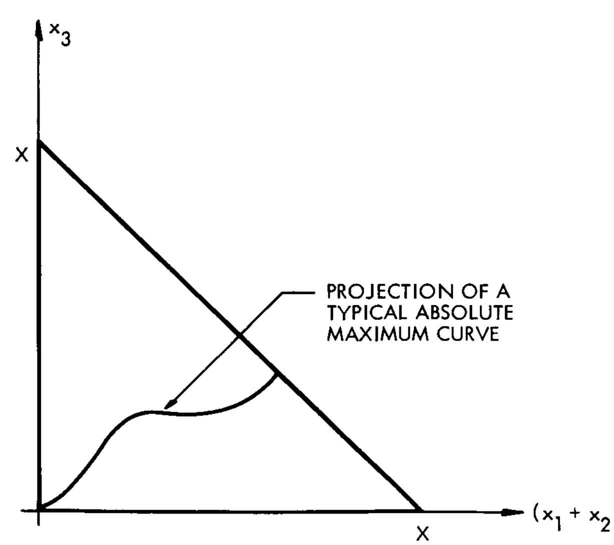

Fig. A-5. Typical projection of the absolute maximum onto the $[(x_1+x_2), x_3]$-plane for the three-activity problem.

two-dimensional search path is indicated in Fig. A-4. By repeating the process for all x in the interval $0 \leqslant x \leqslant X$ a curve such as the one shown in Fig. A-5 is obtained. This curve indicates the division between x_3 and (x_1+x_2) which maximizes $R(x_1, x_2, x_3)$ subject to the constraints. The results of the two-activity allocation (pictured in Fig. A-3) can then be used to determine how the amount of resources represented by the sum (x_1+x_2) is allocated between the first two activities.

It should be clear that as we continue the allocation problem by adding an additional activity we can always reduce the search that is necessary to determine the maximum return to a two-dimensional search by utilizing the results of the previous calculations.

Bellman's technique of dynamic programming has its basic foundation in the type of search which we have just illustrated.

A.4 FORMAL STATEMENT OF THE ALGORITHM

Formally, the search procedure outlined graphically in the preceding section is specified by two sets of functionals; the return functionals $\{r_i(x)\}$ and the allocation or decision functionals $\{X_i(x)\}$. Both of these sets of functionals have members for $i = 2, 3, \ldots, N$, where N is the number of activities that are of interest. These functionals are defined in the following manner:

$$\left.\begin{aligned}
r_2(x) &= \max_{\substack{x_2 \\ 0 \leqslant x_2 \leqslant x \leqslant X}} [a_1(x-x_2)+a_2(x_2)] \\
r_3(x) &= \max_{\substack{x_3 \\ 0 \leqslant x_3 \leqslant x \leqslant X}} [r_2(x-x_3)+a_3(x_3)] \\
&\vdots \\
r_i(x) &= \max_{\substack{x_i \\ 0 \leqslant x_i \leqslant x \leqslant X}} [r_{i-1}(x-x_i)+a_i(x_i)] \\
&\vdots \\
r_N(x) &= \max_{\substack{x_N \\ 0 \leqslant x_N \leqslant x \leqslant X}} [r_{N-1}(x-x_N)+a_N(x_N)]
\end{aligned}\right\} \qquad (A.11)$$

$$\left.\begin{aligned}
X_2(x) &= \text{the value of } x_2 \text{ that maximizes } [a_1(x-x_2)+a_2(x_2)] \\
&\qquad\qquad \text{for the resource } x, \quad 0 \leqslant x \leqslant X; \\
X_3(x) &= \text{the value of } x_3 \text{ that maximizes } [r_2(x-x_3)+a_3(x_3)] \\
&\qquad\qquad \text{for the resource } x, \quad 0 \leqslant x \leqslant X; \\
&\vdots \\
X_N(x) &= \text{the value of } x_N \text{ that maximizes } [r_{N-1}(x-x_N)+a_N(x_N)] \\
&\qquad\qquad \text{for the resource } x, \quad 0 \leqslant x \leqslant X.
\end{aligned}\right\} \qquad (A.12)$$

From an examination of these functionals and from our previous discussion it is clear that $r_N(x)$ enumerates the maximum value of the return, $R(x_1, x_2, \ldots, x_N)$, subject to the constraints for all x in the range $0 \leqslant x \leqslant X$. The specific allocations to the activities are determined from the decision functionals. Suppose that we decide to invest an amount a, $0 \leqslant a \leqslant X$, in the N-activities. Then we invest an amount

$$X_N = X_N(a)$$

in the N^{th} activity, an amount

$$X_{N-1} = X_{N-1}[a-X_N]$$

in the $(N-1)^{th}$ activity, an amount

$$X_{N-2} = X_{N-2}[a-X_N-X_{N-1}]$$

in the $(N-2)^{th}$ activity, and so forth, until

$$X_1 = a - X_N - X_{N-1} - \ldots - X_2$$

remains to be invested in the first activity. These allocations will yield the maximum return.

APPENDIX B

The Quantization Algorithm – A Graphical Search Technique

We want to consider the quantization algorithm as a search technique. We recall from Eq. 4, that the quantization error is given by

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \left[g\left(\xi - y_{i+1}\right) p_x(\xi) \right] \tag{B. 1}$$

where $x_o = X_\ell$ and $x_N = X_u$; X_ℓ and X_u are constants. It should be clear from the similarity between this equation and Eq. A. 1 that the solution to this problem can proceed in a manner similar to the procedure employed in the allocation problem if the term

$$\int_{x_k}^{x_{k+1}} d\xi \left[g\left(\xi - y_{k+1}\right) p_x(\xi) \right] \tag{B. 2}$$

is regarded as the "return" from an activity. There are, however, two differences between these problems. First, the operation to be performed in the quantization problem is minimization rather than maximazation as it was in the allocation problem.

Second, there is a difference in the nature of the surface over which the search is conducted. In the allocation problem we wanted to obtain the maximum return for every allocation x in the interval $0 \leq x \leq X$. Thus, in each of the searches along the subsurfaces of the total return surface it was necessary to search over the entire region $0 \leq x \leq X$. In fact, we observe in the allocation problem that even if we had wanted to obtain the return for a single amount of resources X it would still have been necessary to perform all of the searches over this same interval $0 \leq x \leq X$. This fact is confirmed by a consideration of the nature of the allocation process, (see section A. 4).

In the quantization problem the decisions concerning the placement of the quantizer parameters must be made in exactly the same way as we made the decisions in the allocation problem. Therefore, it will be necessary to permit a variation in the amount of "resources" available for quantization. That is, instead of requiring $x_N = X_u$ we must permit x_N to vary over the region

$$X_\ell \leq x_N \leq X_u.$$

Only by doing this can we obtain the necessary information to make the decisions concerning the optimum parameter placement. Specifically, then, in the quantization problem we shall search along an extended error surface instead of along the error surface, (B. 1).

We are now ready to examine the search technique. In order to simplify our discussion, we make use of the results of section 2. 6 to write Eq. B. 1 as

$$\mathscr{E} = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} d\xi \left[g\left(\xi - y_{i+1}^{*}\right) p_x(\xi) \right]. \tag{B. 3}$$

The y_k^* are defined in section 2.6.

Preceeding as we did in the allocation problem, we begin with two activities; that is, we shall consider two-level quantization. The quantization error for this case is

$$\mathscr{E}(x_1) = \int_{x_0=X_\ell}^{x_1} d\xi \left[g\left(\xi - y_1^*\right) p_x(\xi) \right] + \int_{x_1}^{x_2=X_u} d\xi \left[g\left(\xi - y_2^*\right) p_x(\xi) \right]. \tag{B. 4}$$

The extended error surface over which the search is conducted is specified by

$$\mathscr{E}^*(x_1, x_2) = \int_{x_0=X_\ell}^{x_1} d\xi \left[g\left(\xi - y_1^*\right) p_x(\xi) \right] + \int_{x_1}^{x_2} d\xi \left[g\left(\xi - y_2^*\right) p_x(\xi) \right]. \tag{B. 5}$$

Clearly,

$$\mathscr{E}(x_1) = \mathscr{E}^* (x_1, x_2 = X_u). \tag{B. 6}$$

The minimization for the two-level quantizer will be subject to the constraints expressed by the inequality

$$X_\ell = x_0 \leqslant x_1 \leqslant x_2 \leqslant X_u. \tag{B. 7}$$

Equation B.7 defines the region of variation.

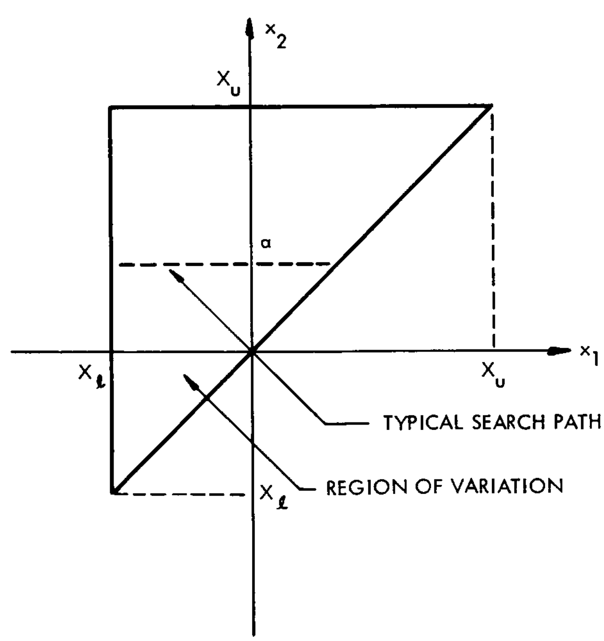

Fig. B-1. A typical search-path in the two-level quantization problem.

One possible search technique is to select a value for x_2, say $x_2 = \alpha$, and search along that portion of this line within the region of variation. This search path and the boundaries to the region of variation are indicated in Fig. B-1. $\mathscr{E}^*(x_1, x_2)$ is plotted along an axis directed out of the page. By examining each of the points on that portion of the search path within the region of variation, the absolute minima of the error for this particular value of $x_2 = \alpha$, is easily obtained. If we repeat this process for all x_2 in the region $X_\ell \leq x_2 \leq X_u$, a curve such as that shown in Fig. B-2. is obtained. This curve indicates the value of x_1 which minimizes the error for any x_2 within the range $X_\ell \leq x_2 \leq X_u$.

Next we consider the problem of minimizing the error involved in three-level quantization. The extended error surface in this case is given by

$$\mathscr{E}^*(x_1, x_2, x_3) = \int_{x_0=X_\ell}^{x_1} d\xi \left[g\left(\xi - y_1^*\right) p_x(\xi) \right] + \int_{x_1}^{x_2} d\xi \left[g\left(\xi - y_2^*\right) p_x(\xi) \right) + \int_{x_2}^{x_3} d\xi \left[g\left(\xi - y_3^*\right) p_x(\xi) \right]. \tag{B.8}$$

The search along this surface will be subject to the usual constraints

$$X_\ell = x_o \leq x_1 \leq x_2 \leq x_3 \leq X_u, \tag{B.9}$$

which define the region of variation.

Upon first inspection it appears that a three-dimensional search will be necessary to determine the parameters that minimize the quantization error. A dimensionality reduction can be achieved in this case just as it was in the allocation case. We observe that since

$$\int_{x_2}^{x_2} d\xi \left[g\left(\xi - y_3^*\right) p_x(\xi) \right] = 0, \tag{B.10}$$

$\mathscr{E}^*(x_1, x_2)$ will be a subsurface of $\mathscr{E}^*(x_1, x_2, x_3)$. Specifically,

$$\mathscr{E}^*(x_1, x_2) = \mathscr{E}^*(x_1, x_2, x_2). \tag{B.11}$$

This result can be employed to reduce the search from three dimensions to two in the way that Eq. A. 10 was used similarly in the allocation problem. The two-dimensional search situation that results upon dimensionality reduction is pictured in Fig. B-3.

If the search indicated in Fig. B-3 is repeated for all x_3 in the interval $X_\ell \leq x_3 \leq X_u$, a curve such as that shown in Fig. B-4 is obtained. This curve indicates the optimum

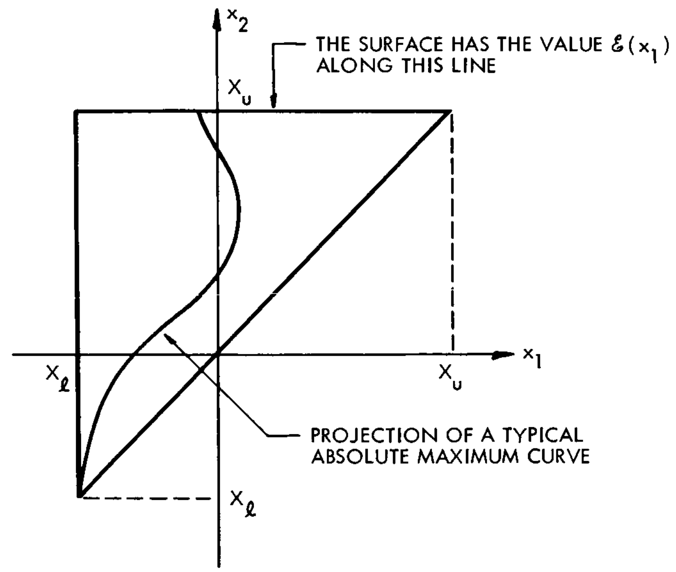

Fig. B-2. Typical projection of the absolute minimum onto the (x_1, x_2)-plane for the two-level quantization problem.

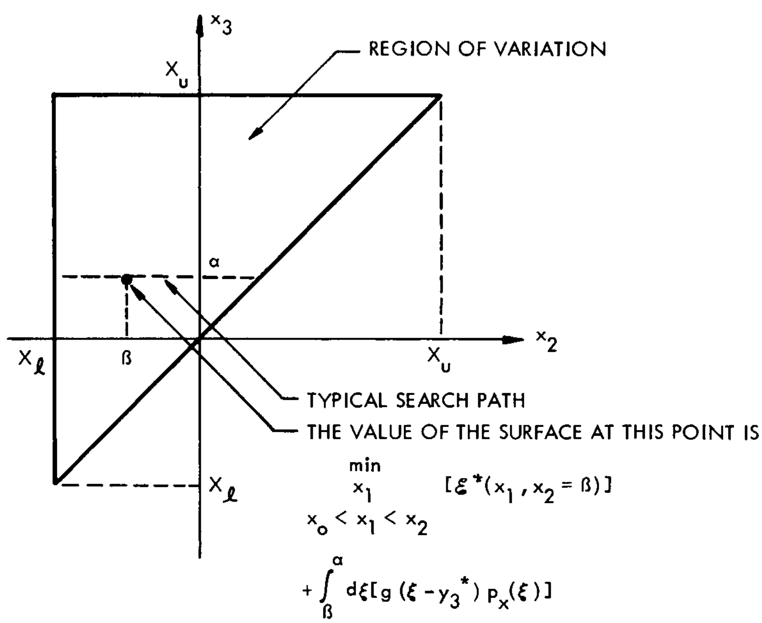

Fig. 3. Typical search path in the three-level quantization problem.

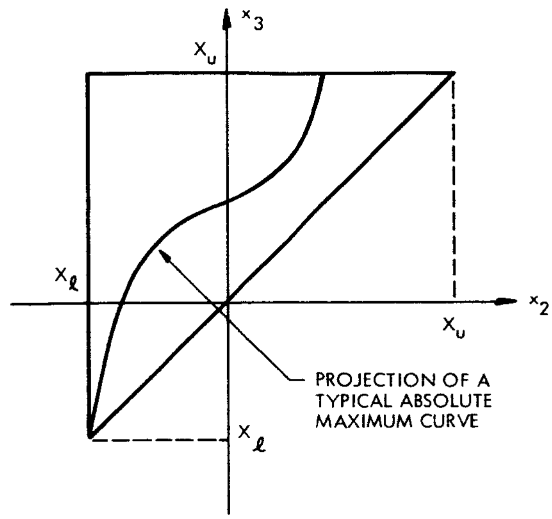

Fig. B-4. Typical projection of the absolute minimum onto the (x_2, x_3)-plane for the three-level quantization problem.

selection for x_2, given a specific x_3. Once this optimum x_2 is obtained, the results of the two-level quantizer search (Fig. B-2) are used to determine the optimum x_1.

Now that we have seen how the search dimensionality is reduced in the three-level quantizer case, it should be evident that this same reduction can be accomplished for each of the remaining levels. Therefore, for the N-level quantizer we have reduced the problem from one involving a search along an (N-1)-dimensional surface to (N-1) searches, each of which is along a two-dimensional surface.

APPENDIX C

Computational Aspects of the Quantization Algorithm

Our objective is to discuss from a computational point of view the quantization algorithm that was presented in Section II. We shall also present the block diagram of a computer program that determines the error and decision functionals. From the discussion presented in section 2.2 we know that the parameters defining the optimum N-level quantizer can be determined from a knowledge of the first N members of these three sets of functionals. In fact, it follows from the nature of these functionals that the parameters defining the optimum K-level, $K \leq N$, quantizer can also be determined from a knowledge of the first N members of these three sets of functionals.

In our original presentation of the quantization algorithm we assumed that the probability density of the input signal x is known for all values of x and that the error and decision functionals are calculated for all values of x within the region $X_\ell \leq x \leq X_u$. When we begin to consider the algorithm from a computational point of view, however, we realize that the calculations necessary to determine the error and decision functionals at every point in the desired interval are too numerous to perform. For this reason, we shall limit our calculations to the determination of the error and decision functionals at M equally spaced grid points covering the interval $X_\ell \leq x \leq X_u$. These grid points will be denoted by the variable ξ_k $k = 1, 2, \ldots, M$. Figure C-1 pictures a typical grid structure. We shall assume that these M grid points are sufficiently dense that as far as the quantizer under design is concerned the error and decision functionals appear to by calculated at every point in the interval $X_\ell \leq x \leq X_u$. The amplitude probability density of the signal will be defined only at the grid points. The value of the amplitude probability density at the grid points is given by

$$P_x(\xi_k) = \int_{(\xi_{k-1}+\xi_k)/2}^{(\xi_k+\xi_{k+1})/2} p_x(\xi)\, d\xi, \qquad k = 1, 2, \ldots, M. \tag{C. 1}$$

Basically, the computational problem is to calculate the error functionals, Eq. 24, since the decision functionals are obtained as an ancillary result of these calculations. A careful examination of the members of Eq. 24 is marked by the appearance of a term of the form

$$\int_{x_k}^{x_{k+1}} d\xi \left[g\left(\xi - y_k^*\right) p_x(\xi) \right] \tag{C. 2}$$

in each member of this set of functionals. Since our method for determining the error functionals is identical to the search technique demonstrated in Appendix B, it will be necessary to calculate every term of the form of Eq. C. 2 (N-1) times in order to

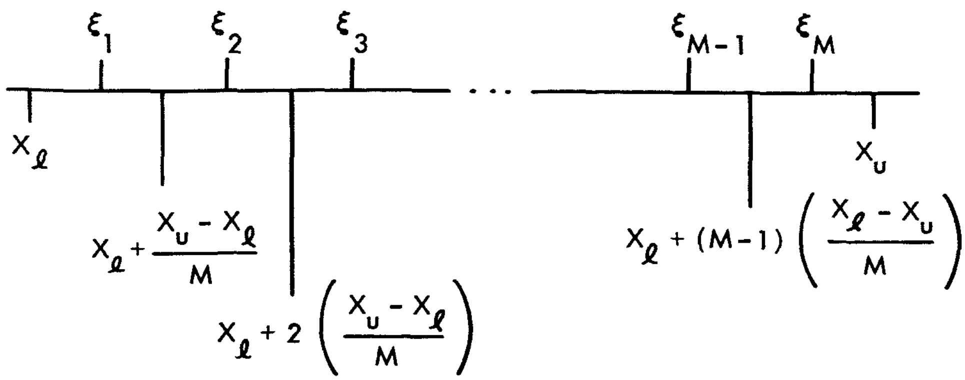

Fig. C-1. Typical grid structure.

determine the first N error functionals. In our discussion we indicated that the functionals will be calculated only at the M grid points. Therefore, there are $\frac{M(M+1)}{2}$ terms of the form of (C. 2). In the interest of computational speed, the values of these $\frac{M(M+1)}{2}$ terms will be calculated once, and then stored for later use in determining the error functionals. The same calculation and storage procedure will be utilized for the y_k^*. By definition

$$\mathrm{TABLE}_{j,k} = \int_{\xi_j}^{\xi_k} d\xi \left[g\left(\xi - y_{j,k}^*\right) p_x(\xi) \right] \tag{C. 3}$$

$$\sum_{\substack{\text{all } i \\ \text{such that} \\ \xi_j \leq \xi_i \leq \xi_k}} g\left(\xi_i - y_{j,k}^*\right) P_x(\xi_i) \tag{C. 4}$$

and

$y_{j,k}^*$ = the value of $y_{j,k}$ that minimizes

$$\int_{\xi_j}^{\xi_k} d\xi [g(\xi - y_{j,k}) p_x(\xi)] \tag{C. 5}$$

= the value of $y_{j,k}$ that minimizes

$$\sum_{\substack{\text{all } i \\ \text{such that} \\ \xi_j \leq \xi_i \leq \xi_k}} g(\xi_i - y_{j,k})\ P_x(\xi_i), \qquad \begin{array}{l} k = 1, 2, \ldots, M \\ j \leq k. \end{array} \tag{C. 6}$$

We have denoted the variable y_k by $y_{j,k}$ in Eqs. C. 3 - C. 6 in order to indentify both end points.

It is interesting to note that in so far as direct computation is concerned, $\text{TABLE}_{j,k}$ and $y^*_{j,k}$ are the only terms dependent on g. This implies that only a small portion of the computer version of the quantization algorithm need be changed in order to change the error-weighting function under consideration.

Now that we have defined the two sets of values, TABLE and y^*, we must consider how they are used to calculate the error and decision functionals. From a comparison of the first error functional, Eq. 24, and TABLE we see that the value of the first error functional at the M grid points is given by

$$\epsilon_1(\xi_k) = \text{TABLE}_{1,k}, \qquad k = 1, 2, \ldots, M. \tag{C. 7}$$

Similarly,

$$Y_1(\xi_k) = y^*_{1,k}, \qquad k = 1, 2, \ldots, M \tag{C. 8}$$

and by using Eq. 8,

$$X_1(\xi_k) = X_\ell, \qquad k = 1, 2, \ldots, M. \tag{C. 9}$$

Now that we have demonstrated that the first error functional can be determined from TABLE, we turn our attention to the second error functional, ϵ_2. Referring to Appendix B, we observe that in order to determine $\epsilon_2(\xi_k)$ [and therefore $Y_2(\xi_k)$ and $X_2(\xi_k)$] we must search the modified error surface along that portion of the line $x_2 = \xi_k$ which is within the region of variation. For $x_2 = \xi_k$ this search will consist of examining the value of the surface at each of the (k) grid points on the line within the region of variation and the boundary point. In order to illustrate this search, let us examine it for the case k = 3. Since k = 3 there are three grid points on the line $x_2 = \xi_3$ within the region of variation. An examination of the grid structure and the error functionals indicates that the first point on this line represents the allocation of that portion of the signal represented by the grid points ξ_1, ξ_2, and ξ_3 to the second quantization interval. The second point on $x_2 = \xi_3$ represents the allocation of ξ_1 to the first quantization interval and ξ_2 and ξ_3 to the second quantization interval. The third point represents the allocation of ξ_1 and ξ_2 to the first-quantization interval and the allocation of ξ_3 to the second interval. The boundary point represents the allocation of ξ_1, ξ_2, ξ_3 to the first quantization interval. From Eqs. 7 and C. 3, the value of the surface at the first point in $\text{TABLE}_{1,3}$; at the second point, $\epsilon_1(\xi_1) + \text{TABLE}_{2,3}$; at the third point, $\epsilon_1(\xi_2) + \text{TABLE}_{3,3}$; and on the boundary $\epsilon_1(\xi_3)$.

The search along this line consists of selecting the minimum of these four values, that is, selecting the minimum of

$$\text{TABLE}_{1,3},$$

$$\epsilon_1(\xi_1) + \text{TABLE}_{2,3},$$

$\epsilon_2(\xi_2) + \text{TABLE}_{3,3},$

$\epsilon_1(\xi_3).$

Assume that the minimum value along this line is at the second point. Then

$$\left.\begin{aligned} \epsilon_2(\xi_3) &= \epsilon_1(\xi_1) + \text{TABLE}_{2,3} \\ Y_2(\xi_3) &= y^*_{2,3}, \\ X_2(\xi_3) &= \xi_1. \end{aligned}\right\} \tag{C. 10}$$

This equation illustrates how each of the points in the error and decision functionals are obtained, once the minimum value on the line of search (in this case $x_2 = \xi_3$) has been obtained. This procedure will be used to search along each of the M lines involved in the determination of the second error and decision functionals, thereby determing $\epsilon_2(\xi_k)$, $Y_2(\xi_k)$, and $X_2(\xi_k)$, $k = 1,2,\ldots,M$.

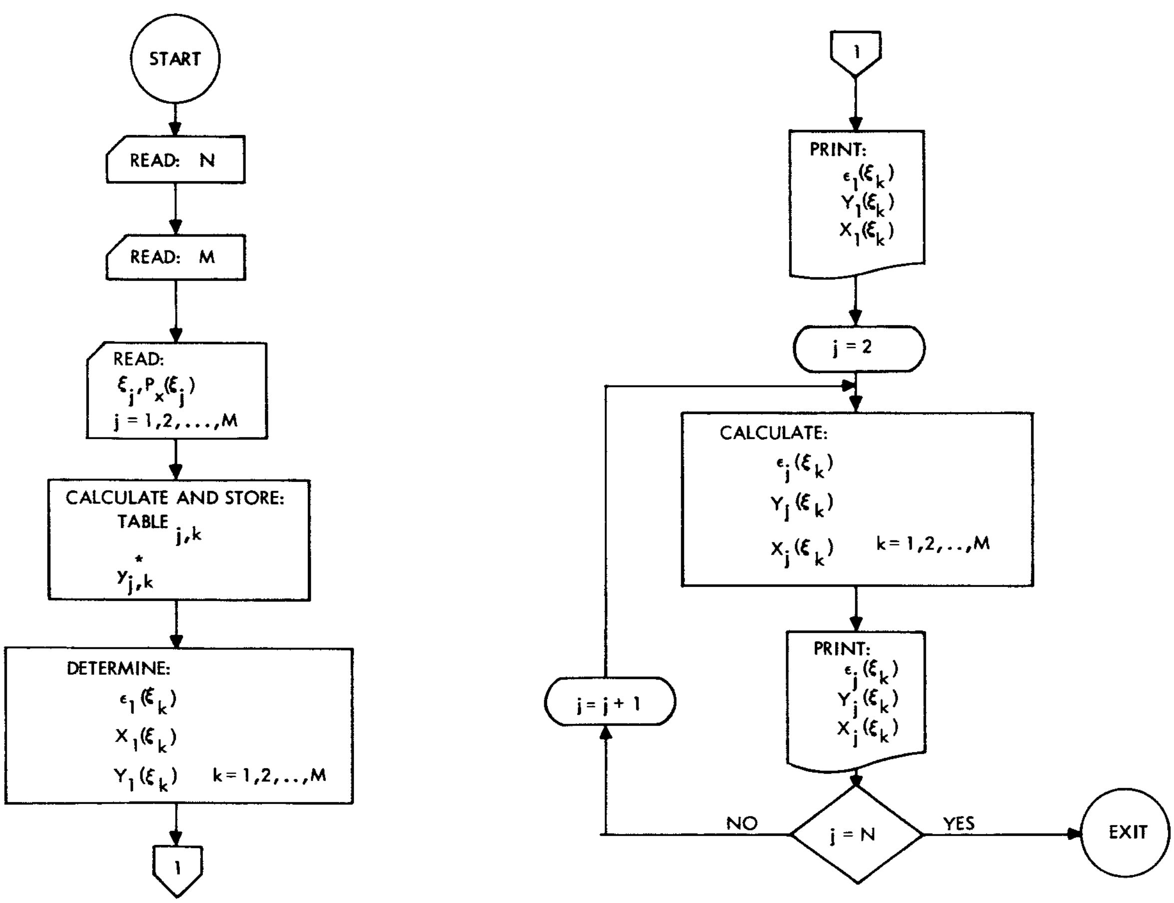

Fig. C-2. Block diagram of the computational version of the quantization algorithm.

There are two special cases of Eq. 42 which should be discussed at this time. First, consider the case in which the minimum is at the first grid point on the line of search. As we have previously shown, this first grid point indicates that all of the signal is being allocated to the second quantization interval. Since by our convention in Eq. 42, $X_2(\xi_k)$ indicates the largest grid point in the first quantization interval, for this special case $X_2(\xi_k)$ will equal X_ℓ. Second, we want to consider the boundary point. This point indicates for the line of search $x_2 = \xi_k$ that the portion of the signal represented by ξ_1, $\xi_1, \xi_2, \ldots, \xi_k$ is all allocated to the first quantization interval. Since there is no allocation to the second quantization interval, $Y_2(\xi_k)$ may be defined to be any convenient value; $X_2(\xi_k)$ will equal ξ_k. In each of these two cases the remaining functional members are determined in the manner indicated by (C. 10).

From Appendix B we recall that the nature of the search necessary to determine each of the remaining error functionals is identical to the search used to determing ϵ_2. Therefore, the methods discussed in connection with ϵ_2 can be applied directly to the calculation of these remaining error functionals.

A block diagram illustrating the basic features of the computational version of the quantization algorithm is presented in Fig. C-2.

Acknowledgment

It is a pleasure to thank Professor Amar G. Bose, who supervised my research, for his many helpful comments and suggestions concerning the research reported here. I also wish to acknowledge the cooperation of the Computation Center, M. I. T., in implementing the quantization algorithms and to thank the Research Laboratory of Electronics, M. I. T., and its sponsors, for extending its support and the use of its facilities.

References

1. W. F. Sheppard, On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale, Proc. London Math. Soc. 29, Part 2, 353-380 (1898).
2. B. Widrow, A Study of Rough Amplitude Quantization by means of Nyquist Sampling Theory, Sc. D. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1956.
3. A. A. Kosyakin, The statistical theory of amplitude quantization, Automatika Telemekh. 22, 722-729 (1961).
4. W. M. Goodall, Telephony by pulse-code modulation, Bell System Tech. J. 26, 395-409 (1947).
5. W. R. Bennett, Spectra of quantized signals, Bell System Tech. J. 27, 446-472 (1948).
6. W. R. Bennett, Statistics of regenerative digital transmission, Bell System Tech. J. 37, 1501-1542 (1958).
7. A. I. Velichkin, Correlation function and spectral density of a quantized process, Telecommunications and Radio Engineering, Part II, No. 7, pp. 70-77, 1962.
8. D. S. Ruchkin, Optimal Reconstruction of Sampled and Quantized Stochastic Signals, Engr. D. Thesis, Yale University, 1960.
9. B. Smith, Instantaneous companding of quantized signals, Bell System Tech. J. 36, 653-709 (1957).
10. I. A. Lozovoy, Regarding the computation of the characteristics of compression in systems with pulse-code modulation, Telecommunications, No. 10, pp. 18-25, 1961.
11. H. Mann, H. M. Straube, and C. P. Villars, A companded coder for an experimental PCM terminal, Bell System Tech. J. 41, 173-226 (1962).
12. C. G. Davis, An experimental pulse-code modulation system for short-haul systems, Bell System Tech. J. 41, 1-24 (1962).
13. R. H. Shemmum and J. R. Gray, Performance limitations of a practical PCM terminal, Bell System Tech. J. 41, 143-172 (1962).
14. M. J. Wiggins and R. A. Branham, Reduction in quantizing levels for digital voice transmission, 1963 IEEE International Convention Record, Part 8, pp. 282-288.
15. R. F. Purton, A survey of telephone speech-signal statistics and their significance in the choice of a PCM companding law, Proc. Inst. Elec. Engrs. (London) 109B, 60-66 (1962).
16. J. Katzenelson, A Note on Errors Introduced by Combined Sampling and Quantization, Technical Memorandum ESL-TM-101, Electronics Systems Laboratory, Massachusetts Institute of Technology, March 1961.
17. J. J. Stiffler, On the determination of quantization levels in a data sampling system, JPL Space Programs Summary No. 37-18, Vol. 4, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, December 31, 1962.
18. D. N. Graham, Two-dimensional Filtering to Reduce the Effect of Quantizing Noise in Television, S. M. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, 1962.
19. H. A. Spang III, Quantizing Noise Reduction, Report No. 62-RL-(2999E), General Electric Research Laboratory, Schenectady, New York, April 1962.
20. H. A. Spang III and P. M. Schulthesis, Reduction of quantizing noise by use of feedback, IRE Trans., Vol. CS-10, pp. 373-380, December 1962.
21. E. G. Kimme and F. F. Kuo, Synthesis of optimal filters for a feedback quantization system, 1963 IEEE International Convention Record, Part 2, pp. 16-26.

22. C. C. Cutler, Transmission Systems Employing Quantization, Patent No. 2,927,962, March 8, 1960 (filed April 26, 1954).

23. G. G. Furman, Removing the Noise from the Quantization Process by Dithering: Linearization, Memorandum RM-3271-PR, The Rand Corporation, Santa Monica, California, February 1963.

24. G. G. Furman, Improving the Quantization of Random Signals by Dithering, Memorandum RM-3504-PR, The Rand Corporation, Santa Monica, California, May 1963.

25. L. G. Roberts, PCM Television Bandwidth Reduction Using Pseudo-Random Noise, S.M. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1961.

26. J. Max, Quantizing for minimum distortion, IRE Trans., Vol. IT-6, pp. 7-12, March 1960.

27. S. P. Lloyd, Least Squares Quantization in PCM (unpublished manuscript, Bell Telephone Laboratories, Inc., 1958) reported by Ruchkin[6] and Spang.[15]

28. V. A. Gamash, Quantization of signals with non-uniform steps, Electrosvyaz 10, 11-13 (October 1957).

29. L. I. Bluestein, A Hierarchy of Quantizers, Ph.D. Thesis, Department of Electrical Engineering, Columbia University, New York, 1962.

30. J. T. Tou, Optimum Design of Digital Control Systems (Academic Press, New York, 1963), pp. 103-169.

31. G. M. Roe, Personal correspondence, July 1963.

32. G. H. Myers, Quantization of a signal plus random noise, IRE Trans., Vol. I-5, pp. 181-186, June 1956.

33. I. B. Stiglitz, Level Transformation and Multibit Dimus, Report 835-223-15, General Atronics Corporation, Philadelphia, Pennsylvania, April 1961.

34. L. I. Bluestein and R. J. Schwarz, Optimum zero memory filters, IRE Trans., Vol. IT-8, pp. 337-342, October 1962.

35. A. M. Kuperman, Application of the Theory of Statistical Decisions to some Level Quantization Problems, Automation and Remote Control, Vol. 24, No. 12, pp. 1538-1544, December 1963 (translated from Automatika i Telemekhanika, Vol. 24, No. 12, pp. 1685-1691, translation published May 1964).

36. T. M. Apostol, Calculus, Vol. I; Introduction, with Vectors and Analytic Geometry (Blaisdell Publishing Company, New York, 1961), pp. 375 ff.

37. R. Bellman, Dynamic Programming (Princeton University Press, Princeton, N.J., 1957).

38. R. Bellman and J. Dreyfus, Applied Dynamic Programming (Princeton University Press, Princeton, N.J., 1962).

39. R. Bellman and B. Kotkin, On the Approximation of Curves by Linear Segments Using Dynamic Programming – II, Memorandum RM-2978-PR, The Rand Corporation, Santa Monica, California, February 1962.

40. I. S. Sokolnikoff and R. M. Redheffer, Mathematics of Physics and Modern Engineering (McGraw-Hill Book Company, Inc., New York, 1958), pp. 247-248.

41. P. E. Fleischer, Sufficient conditions for achieving minimum distortion in a quantizer, 1964 IEEE International Convention Record, Part 1, pp. 104-111.

42. R. Bellman, Introduction to Matrix Analysis (McGraw-Hill Book Company, Inc., New York, 1960), pp. 294-295.

43. H. Cramér, Mathematical Methods of Statistics (Princeton University Press, Princeton, N.J., 1945), pp. 271-272.

44. D. A. Chesler, Nonlinear Systems with Gaussian Inputs, Technical Report 366, Research Laboratory of Electronics, M. I. T., Cambridge, Mass., February 15, 1960, pp. 43-45.

45. J. D. Egan, Articulation Testing Methods II, Harvard Psycho-Acoustic Laboratory Report, OSRD No. 3802, Harvard University, November 1944.

46. W. Ryan, A Method for Optimum Speech Quantization, S. M. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1963.

47. W. R. Davenport, Jr., A Study of Speech Probability Distributions, Technical Report 148, Research Laboratory of Electronics, M. I. T., Cambridge, Mass., August 25, 1950.

48. G. F. Crimi, The Effect of Pre- and Post-emphasis Filtering on an Optimum Quantization System, S. M. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1964.

49. V. R. Algazi, A Study of the Performance of Linear and Nonlinear Filters, Sc. D. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1963.

50. A. V. Balakrishnan, On a characterization of processes for which optimal mean-square systems are of a specified form, IRE Trans., Vol. IT-6, pp. 490-500, September 1960.

51. S. E. Dreyfus, Dynamic programming and the calculus of variations, J. Math. Anal. Applicat. 1, pp. 228-239, 1960.

JOINT SERVICES DISTRIBUTION LIST

Department of Defense

Dr Edward M. Reilley
Asst Director (Research)
Ofc of Defense Res & Eng
Department of Defense
Washington, D.C. 20301

Dr James A. Ward
Office of Deputy Director (Research and Information Rm 3D1037
Department of Defense
The Pentagon
Washington, D.C. 20301

Director
Advanced Research Projects Agency
Department of Defense
Washington, D.C. 20301

Mr Charles Yost, Director
For Materials Sciences
Advanced Research Projects Agency
Department of Defense
Washington, D.C. 20301

Defense Documentation Center
Attn: TISIA
Cameron Station, Bldg 5
Alexandria, Virginia 22314

Director
National Security Agency
Attn: C3/TDL
Fort George G. Meade, Maryland 20755

Department of the Army

Chief of Research and Development
Headquarters, Department of the Army
Attn: Physical Sciences Division P&E
Washington, D.C. 20310

Research Plans Office
U.S. Army Research Office
3045 Columbia Pike
Arlington, Virginia 22204

Commanding Officer
Foreign Service & Technology Center
Arlington Hall
Arlington, Virginia

Commanding General
U.S. Army Materiel Command
Attn: AMCRD-RS-PE-E
Washington, D.C. 20315

Commanding General
U.S. Army Strategic Communications Command
Washington, D.C. 20315

Commanding General
U.S. Army Materials Research Agency
Watertown Arsenal
Watertown, Massachusetts 02172

Commanding Officer
U.S. Army Ballistics Research Laboratory
Attn: V.W. Richards
Aberdeen Proving Ground
Aberdeen, Maryland 21005

Commandant
U.S. Army Air Defense School
Attn: Missile Sciences Division, C&S Dept.
P.O. Box 9390
Fort Bliss, Texas 79916

Commanding General
U.S. Army Missile Command
Attn: Technical Library
Redstone Arsenal, Alabama 35809

Commanding General
Frankford Arsenal
Attn: SMUFA-1310 (Dr Sidney Ross)
Philadelphia, Pennsylvania 19137

U.S. Army Munitions Command
Attn: Technical Information Branch
Picatinney Arsenal
Dover, New Jersey 07801

Commanding Officer
Harry Diamond Laboratories
Attn: Mr Berthold Altman
Connecticut Avenue and Van Ness Street N.W.
Washington, D.C. 20438

Commanding Officer
Harry Diamond Laboratories
Attn: Dr R.T. Young
Electron Tubes Division
Connecticut Avenue and Van Ness Street N.W.
Washington, D.C. 20438

Commanding Officer
U.S. Army Security Agency
Arlington Hall
Arlington, Virginia 22212

Commanding Officer
U.S. Limited War Laboratory
Attn: Technical Director
Aberdeen Proving Ground
Aberdeen, Maryland 21005

Commanding Officer
Human Engineering Laboratories
Aberdeen Proving Ground
Maryland 21005

Director
U.S. Army Engineer Geodesy,
Intelligence and Mapping
Research and Development Agency
Fort Belvoir, Virginia 22060

Commandant
U.S. Army Command and General Staff
College
Attn: Secretary
Fort Leavenworth, Kansas 66207

Dr. H. Robl, Deputy Director
U.S. Army Research Office (Durham)
P.O. Box CM, Duke Station
Durham, North Carolina 27706

Commanding Officer
U.S. Army Research Office (Durham)
Attn: CRD-AA-IP (Richard O. Ulsh)
P.O. Box CM, Duke Station
Durham, North Carolina 27706

Commanding General
U.S. Army Electronics Command
Attn: AMSEL-SC
Fort Monmouth, New Jersey 07703

Director
U.S. Army Electronics Laboratories
Attn: Dr S. Benedict Levin, Director
Institute for Exploratory Research
Fort Monmouth, New Jersey 07703

Director
U.S. Army Electronics Laboratories
Attn: Mr Robert O. Parker, Executive
Secretary JSTAC (AMSEL-RD-X)
Fort Monmouth, New Jersey 07703

Superintendent
U.S. Army Military Academy
West Point, New York 10996

The Walter Reed Institute of Research
Walter Reed Army Medical Center
Washington, D.C. 20012

Director
U.S. Army Electronics Laboratories
Fort Monmouth, New Jersey 07703

Attn: AMSEL-RD-DR	NE	SS	
X	NO	PE	
XE	NP	PR	
XC	SA	PF	
XS	SE	GF	
NR	SR	ADT	
		FU#1	

Commanding Officer
U.S. Army Electronics R&D Activity
Fort Huachuca, Arizona 85163

Commanding Officer
U.S. Army Engineers R&D Laboratory
Attn: STINFO Branch
Fort Belvoir, Virginia 22060

Commanding Officer
U.S. Army Electronics R&D Activity
White Sands Missile Range
New Mexico 88002

Director
Human Resources Research Office
The George Washington University
300 N. Washington Street
Alexandria, Virginia 22314

Commanding Officer
U.S. Army Personnel Research Office
Washington, D.C.

Commanding Officer
U.S. Army Medical Research Laboratory
Fort Knox, Kentucky

Department of the Air Force

Director
Air University Library
Maxwell A.F. Base, Alabama

Commander
Air Force Office of Scientific Research
Washington 25, D.C.
Attn:-SREE

Department of The Air Force
Headquarters-United States Air Force
Washington 25, D.C.
Attn: AFTAC/TD-1

Dr. Harvey E. Savely, SRL
Air Force Office of Sci. Res.
Office of Aerospace Research, USAF
Washington 25, D.C.

Mr. C.N. Hasert
Scientific Advisory Board
Hq, USAF
Washington 25, D.C.

APGC (PGBAP-1)
Elgin Air Force Base
Florida 32542

AFETR
(AFETR Tech. Library MU-135)
Patrick Air Force Base
Cocoa, Florida

Air Force Cambridge Res. Lab.
L.G. Hanscom Field
Bedford, Massachusetts 01731
Attn: CRDM, Mr. Herskovitz

Commander, AFCRL
Attn: C.P. Smith (CRBS)
L.G. Hanscom Field
Bedford, Massachusetts

Dr. L. C. Block
AFCRL (CROV)
L. G. Hanscom Field
Bedford, Massachusetts

AFCRL
Office of Aerospace Res., USAF
Bedford, Mass.
Attn: CRDA

Mr. Rocco H. Urbano, Chief
AFCRL, Appl. Math. Branch
Data Sciences Laboratory
Laurence G. Hanscom Field
Bedford, Massachusetts

AFCRL (CRFE-Dr. Nicholas Yannoni)
L.G. Hanscom Field
Bedford, Massachusetts

S. H. Sternick
Aerospace Comm. - Attn: ESNC
Waltham Federal Center
424 Trapelo Road
Waltham, Massachusetts 02154

Rome Air Dev. Center (RAWL, H.Webb)
Griffiss Air Force Base
New York 13442

Systems Engineering Group
Deputy for Systems Eng'g., SEPRR
Directorate of Tech. Pubs. and Specs.
Wright-Patterson AFB, OHIO 45433

Aeronautical Systems Division
Attn: ASRPE, Mr. Robt. Cooper
Wright-Patterson AFC, Ohio 45433

Aeronautical Systems Division
Attn: ASRPP-20 (Mr. Don R. Warnock)
Wright-Patterson AFB, Ohio 45433

AFAL
AVR (L)
Wright-Patterson AFB
Ohio 45433

Dr. H. H. Kurzweg
Director Research - OART
NASA
Washington, D.C. 20546

Systems Engineering Group (RTD)
Attn: SEPIR
Wright-Patterson AFB, Ohio 45433

AFAL (AVTE)
Wright-Patterson AFB
Ohio 45433

Mr. Roland Chase
National Aeronautics & Space Administration
1512 H Street, N.W.
Washington 25, D.C.

Professor Arwin Dougal
University of Texas
EE Department
Austin, Texas

Honorable Alexander H. Flax
Asst Secretary of the Air Force (R&D)
Office of the Secretary of the Air Force
Washington 25, D.C.

Professor Nicholas George
California Institute of Technology
EE Department
Pasadena, California

Dr. Lowell M. Hollingsworth
AFCRL
L.G. Hanscom Field
Bedford, Massachusetts

Dr. Zohrab Kaprielian
University of Southern California
University Park
Los Angeles 7, California

JOINT SERVICES DISTRIBUTION LIST (continued)

Dr. John M. Ide
National Science Foundation
Washington 25, D.C.

Lt Col Edwin M. Myers
Headquarters USAF (AFRDR)
Washington 25, D.C.

Professor Wm. H. Radford
Director, Lincoln Laboratories
Lexington, Massachusetts

Brig Gen B.G. Holzman, USAF (Ret.)
Electronics Research Center, NASA
30 Memorial Drive
Cambridge, Mass.

Dr. R. L. Sproull
Director, Advanced Research Projects Agency
Washington 25, D.C.

Brigadier General J. T. Stewart
Director of Science & Technology
Deputy Chief of Staff (R&D)
USAF
Washington 25, D.C.

Mr. James Tippett
National Security Agency
Fort Meade, Maryland

Dr. H. Harrison
NASA (Code RRE)
Fourth and Independence Streets
Washington, D.C. 20546

AEC
Civ of Tech Info Ext
P.O. Box 62
Oak Ridge, Tenn.

AFRST (SC/EN)
Lt Col L. Stone
Rm 4C 341
The Pentagon
Washington, D.C. 20301

U.S. Atomic Energy Commission
Library
Gaithersburg, Md. 20760

ARL (ARD/Col R.E. Fontana)
Wright-Patterson AFB,
Ohio 45433

Office of Research Analyses
Attn: Col K.W. Gallup
Holloman AFB, NMex 88330

AFCRL (CRXL)
L.G. Hanscom Fld
Bedford, Mass 01731

Frank J. Seiler Rsch Lab
Library
USAF Academy, Colo 80840

ARL (AROL)
Wright-Patterson AFB,
Ohio 45433

Office of Research Analyses
Library
Holloman AFB, NMex 88330

LOOAR (Library)
AF Unit Post Office
Los Angeles, Calif 90045

Churchill Research Range
Library
Fort Churchill
Manitoba, Canada

Los Alamos Scientific Lab
Attn: Technical Library
Los Alamos, NMex 87544

Battelle Memorial Institute
Technical Library
505 King Avenue
Columbus, Ohio 43201

John Crerar Library
35 West 33rd St.
Chicago, Ill.

Linda Hall Library
5109 Cherry St.
Kansas City, Mo.

National Science Foundation
Library
1951 Constitution Ave., N.W.
Washington, D.C. 20550

Johns Hopkins University
Applied Physics Lab Library
White Oak
Silver Spring, Md. 20910

Stanford Research Institute
Library
820 Mission St.
South Pasadena, Calif. 91030

Southwest Research Institute
Library
8500 Culebra Road
San Antonio, Texas

ARPA, Tech Info Office
The Pentagon
Washington, D.C. 20301

DDR&E (Tech Library)
Rm 3C 128
The Pentagon
Washington, D.C. 20301

Industrial College of the
Armed Forces
Attn: Library
Washington, D.C.

AFIT (MCLI)
Tech Library
Wright-Patterson AFB
Ohio 45433

AUL 3T-9663
Maxwell AFB, Ala 36112

USAFA (DLIB)
USAF Academy, Colorado 80840

AFSC (Tech Library)
Andrews AFB
Washington, D.C. 20331

ASD (Tech Library)
Wright-Patterson, AFB
Ohio 45433

BSD (Tech Library)
Norton AFB, Calif 92409

ESD (ESTI)
L. G. Hanscom Field, F172
Bedford, Mass 01731

RTD (Tech Library)
Bolling AFC, D.C. 20332

AFFTC (Tech Library)
Edwards AFB, Calif 93523

AFMDC (Tech Library)
Holloman AFB, NMex 88330

AFWL (WLIL, Tech Library)
Kirtland AFB, NMex 87117

APGC (Tech Library)
Eglin AFB, Fla 32542

AEDC (Tech Library)
Arnold AFS, Tenn 37389

RADC (Tech Library)
Griffiss AFB, N.Y. 13442

Director
National Aeronautical Establishment
Ottawa, Ontario, Canada

CIA
OCR/LY/IAS
IH 129 Hq
Washington, D.C. 20505

National Defense Library
Headquarters
Ottawa, Ontario, Canada

Technical Library
White Sands Missile Range
NMex 88002

NASA/AFSS/1 FOB6
Tech Library, Rm 60084
Washington, D.C. 20546

Space Systems Division
Los Angeles Air Force Station
Air Force Unit Post Office
Los Angeles, California 90045
Attn: SSSD

U.S. Regional Science Office/LAOAR
U.S. Embassy
APO 676
New York, N.Y.

Ames Rsch Center (NASA)
Technical Library
Moffett Field, Calif 94035

High Speed Flight Center (NASA)
Technical Library
Edwards AFB, Calif 93523

Goddard Space Flight Center (NASA)
Greenbelt, Md. 20771

Geo. C. Marshall Space Flight
Center (NASA)
Redstone Arsenal, Ala 35808

Lewis Research Center (NASA)
Technical Library
21000 Brookpark Road
Cleveland, Ohio

Aerospace Corp (Tech Library)
P.O. Box 95085
Los Angeles, Calif 90045

Rand Corporation
1700 Main St.
Santa Monica, Calif 90401

Carnegie Institute of Technology
Science & Engineering Hunt Library
Schenley Park
Pittsburgh, Pa. 15213

California Institute of Technology
Aeronautics Library
1201 East Calif St.
Pasadena 4, Calif

AVCO Research Lab
Library
2385 Revere Beach Parkway
Everett, Mass 02149

Dr. G. E. Knausenberger
c/o Hq. Co. Munich Post
APO 09407
New York, N. Y.

Commander
Space Systems Division (AFSC)
Office of the Scientific Director
Inglewood, California

Commander
Aerospace Systems Division
AFSC
Office of the Scientific Director
Wright-Patterson AFB, Ohio

Commander
Aerospace Research Laboratories (OAR)
Office of the Scientific Director
Wright-Patterson AFB, Ohio

Commander
Air Force Cambridge Research Laboratories
Office of the Scientific Director
L. G. Hanscom Field
Bedford, Massachusetts

Commander
Air Force Systems Command
Office of the Chief Scientist
Andrews AFB, Maryland

Commander
Research & Technology Division
AFSC
Office of the Scientific Director
Bolling AFB 25, D.C.

Commander
Rome Air Development Center
AFSC
Office of the Scientific Director
Griffiss AFB, Rome, New York

Department of the Navy

Dr. Arnold Shostak, Code 427
Head, Electronics Branch
Physical Sciences Division
Department of the Navy
Office of Naval Research
Washington, D.C. 20360

Chief of Naval Research, Code 427
Department of the Navy
Washington, D. C. 20360

Chief, Bureau of Weapons
Department of the Navy
Washington, D.C. 20360

Chief, Bureau of Ships
Department of the Navy
Washington, D. C. 20360
Attn: Code 680

Commander
U.S. Naval Air Development Center
Johnsville, Pennsylvania
Attn: NADC Library

JOINT SERVICES DISTRIBUTION LIST (continued)

Library
U.S. Navy Electronics Laboratory
San Diego, California 92152

Commanding Officer
U.S. Navy Underwater Sound Laboratory
Ft Trumbull
New London, Connecticut

Director
Naval Research Laboratory
Washington, D.C. 20390

Commanding Officer
Office of Naval Research Branch Office
Navy 100, Fleet P.O. Box 39
New York, New York

Chief of Naval Operations
Pentagon OP 07T
Washington, D.C.

Commanding Officer
Officer of Naval Research Branch Office
495 Summer Street
Boston, Massachusetts 02110

Commander
Naval Ordnance Laboratory
White Oak, Maryland
Attn: Technical Library

U.S. Navy Post Graduate School
Monterey, California
Attn: Electrical Engineering Department